การศึกษาเปรียบเทียบการจัดกลุ่มข้อมูลด้วยอัลกอลิทึมแบบต่าง ๆ

The COMPARISON STUDIES OF VARIOUS CLUSTERING ALGORITHMS

โดย

ณัฐจักร์ เครือจิรายุส์

อาจารย์ที่ปรึกษา

รศ.ดร. อาริต ธรรมโน

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2550

THE COMPARISON STUDIES OF VARIOUS CLUSTERING ALGORITHMS

NUTTAJAK KRUAJIRAYU

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/2007

ใบรับรองปริญญานิพนธ์ ประจำปีการศึกษา 2550

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง การศึกษาเปรียบเทียบการจัดกลุ่มข้อมูลด้วยอัลกอลิทึมแบบต่าง ๆ

The Comparison Studies of Various Clustering Algorithms

ผู้จัดทำ

นายณัฐจักร์ เครือจิรายุส์ **รหัสประจำตัว** **47070027**

.....**อาจารย์ที่ปรึกษา**

(รศ.ดร. อาริต ธรรมโน)

หัวข้อวิทยานิพนธ์	การศึกษาเปรียบเทียบการจัดกลุ่มข้อมูลด้วยอัลกอลิทึมแบบต่าง ๆ
นักศึกษา	นายณัฐจักร์ เครือจิรายุส์
รหัสนักศึกษา	47070027
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
ปีการศึกษา	2550
อาจารย์ที่ปรึกษา	รศ.ดร.อาริต ธรรมโน

บทคัดย่อ

การทำดาด้าไมนิ่งถูกนำมาใช้วิเคราะห์ข้อมูล และจัดกลุ่มข้อมูลเพื่อให้ความรู้ใหม่ๆที่ซ่อนอยู่ในฐานข้อมูล ปัจจุบันข้อมูลเหล่านี้มีจำนวนมากขึ้น ทำให้เราสามารถนำข้อมูลเหล่านี้ไปใช้ประโยชน์ได้ดียิ่งขึ้น ดังนั้นในปัจจุบันจึงได้มีการคิดค้นวิธีการต่างๆในการนำข้อมูลเหล่านี้มาใช้ให้เกิดประโยชน์ ซึ่งผลลัพธ์ที่ได้จากการวิเคราะห์จะถูกนำมาใช้ในการวางแผนกลยุทธ์ทางการตลาด

โมเดลที่ใช้ในการทำดาด้าไมนิ่งนั้นมีมากมาย แต่ที่สนใจนำมาศึกษาคือโมเดลที่ใช้ในการจัดกลุ่มข้อมูล ซึ่งเลือกใช้อัลกอริทึม K-means, PAM, CLARA, CLARANS, Fuzzy C-means, Self organizing map (SOM)

Thesis Title	The Comparison Studies of Various Clustering Algorithms.
Student	Mr. Nuttajak Kruajirayu
Student ID.	47070027
Degree	Bachelor of Science
Major	Information Technology
Academic Year	2007
Advisor	Assoc. Prof. Dr. Arit Thammano

ABSTRACT

Data mining is used for data analysis and supports knowledge discover by finding hiden data in databases and these data are clustered or grouped based. In present there are large amounts of data which need to bring to analysis for the most useful. We use that result to do marketing strategy and make decision with business plan.

There are many data mining model to use but we almost interesting and often using is clustering model which have K-means algorithm, PAM algorithm, CLARA algorithm, CLARANS algorithm, Fuzzy C-means algorithm, Self organizing map algorithm (SOM)

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบัน การทำงานของแต่ละองค์กรจะมีการเก็บรวบรวมข้อมูลเป็นจำนวนมาก แต่ในหลายองค์กรไม่สามารถที่จะนำเอาข้อมูลที่มีประโยชน์เหล่านี้มาใช้งานอย่างคุ้มค่า เนื่องจากข้อมูลในฐานข้อมูลที่นำมาใช้วิเคราะห์และวางแผนทางกลยุทธ์นับวันจะมีจำนวนมากขึ้นเรื่อยๆจนเกินความสามารถของคนที่จะทำการวิเคราะห์ข้อมูลเหล่านี้ด้วยตนเอง ดังนั้นจึงมีการนำวิธีการต่างๆ เพื่อมาใช้วิเคราะห์ข้อมูลที่มีอยู่เพื่อให้เกิดประโยชน์สูงสุด วิธีการหนึ่งที่นิยมใช้ คือ Data Mining

Data Mining เป็นวิธีการค้นหาความรู้จากข้อมูลขนาดใหญ่ ซึ่งเป็นความรู้ที่ผู้ใช้ไม่ทราบมาก่อน โดยค้นหาความสัมพันธ์ รูปแบบหรือโครงสร้างของข้อมูล การจัดกลุ่มข้อมูลเป็นรูปแบบหนึ่งในการทำดาต้าไมนิ่ง คือ การรวมกลุ่มของสิ่งต่างๆที่คล้ายคลึงกันจากกลุ่มใหญ่ให้เป็นกลุ่มย่อยหรือคลัสเตอร์ (Cluster) เรียกว่า คลัสเตอร์ริ่ง ซึ่งการรวมตัวจะไม่พึ่งพาอาศัยการกำหนดหมวดหมู่ล่วงหน้า และไม่ใช้ตัวอย่าง ข้อมูลจะรวมตัวกันบนพื้นฐานของความคล้ายในตัวเอง โดยอัลกอริทึมในการทำการจัดกลุ่มมีอยู่หลายอัลกอริทึม แต่ละอัลกอริทึมก็มีวัตถุประสงค์และประสิทธิภาพของการวิเคราะห์ที่ต่างกัน จึงได้ทำการศึกษาหลักการและทฤษฎีของการจัดกลุ่มข้อมูล (Clustering) เพื่อให้สามารถเลือกใช้อัลกอริทึมแบบต่างๆ ตรงตามความต้องการ และเกิดประสิทธิภาพในการนำข้อมูลไปใช้ประโยชน์ได้มากขึ้น

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1.2.1 เพื่อศึกษาวิธีการและขั้นตอนการทำงานของการจัดกลุ่มข้อมูล

1.2.2 เพื่อศึกษากระบวนการในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมแบบต่างๆ

1.2.3 เพื่อนำเอาความรู้ในกระบวนการจัดกลุ่มด้วยอัลกอริทึมต่างๆไปประยุกต์ใช้ในแอพพลิเคชั่น

1.2.4 เพื่อนำข้อมูลที่ได้จากการจัดกลุ่มไปใช้ให้เกิดประโยชน์และตรงตามความต้องการ

1.3 ขอบเขตการศึกษา

1.3.1 ทำการออกแบบและพัฒนาโปรแกรม

1.3.2 ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลชนิดจำนวนจริง

1.3.3 ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลที่เป็นข้อมูลแบบ offline

1.3.4 โครงงานนี้ขอกล่าวเฉพาะอัลกอลิทึมที่เป็นการจัดกลุ่มโดยใช้วิธีแบ่งกลุ่ม

1.4 ขั้นตอนของการศึกษา

1.4.1 ทำการศึกษาหลักการ ทฤษฎี และอัลกอริทึมต่างๆ ของการจัดกลุ่มข้อมูลแบบแบ่งกลุ่ม (Partitioning Methods)

1.4.2 เลือกข้อมูลและเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสม

1.4.3 ออกแบบและพัฒนาระบบเพื่อวิเคราะห์ข้อมูล

1.4.4 ทดสอบระบบ

1.4.5 ประเมินผลและสรุป

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 เพื่อได้เรียนรู้ทฤษฎี เทคนิคและวิธีการของการจัดกลุ่มข้อมูล

1.5.2 เพื่อนำเอาความรู้ที่มีไปประยุกต์ใช้ในการพัฒนาโปรแกรมเพื่อทำการจัดกลุ่มข้อมูลต่างๆ

1.5.3 เพื่อนำเอาข้อมูลที่มีอยู่มาวิเคราะห์และนำไปใช้ให้เกิดประโยชน์มากขึ้น

1.5.4 เพื่อให้ผู้ใช้สามารถเลือกใช้อัลกอริทึมต่างๆในการจัดกลุ่มข้อมูลตรงตามความต้องการได้

บทที่ 2
ทฤษฎีและหลักการที่เกี่ยวข้อง

2.1 การจัดกลุ่ม (Clustering)

การจัดกลุ่ม (Clustering) คือ การจัดกลุ่มข้อมูลที่มีความเหมือนหรือคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกันและข้อมูลที่มีความแตกต่างกันให้อยู่ต่างกลุ่มกัน โดยไม่มีการจัดกลุ่มข้อมูลตัวอย่างไว้ล่วงหน้า ซึ่งการจัดกลุ่มข้อมูลจะพิจารณาจากค่าความเหมือนหรือคล้ายคลึงกันของข้อมูล

2.2 การเตรียมข้อมูลให้เหมาะสมสำหรับใช้ในการวิเคราะห์การจัดกลุ่ม

ข้อมูลหรือข้อมูลดิบ มาจากข้อเท็จจริงหรือเหตุการณ์ต่างๆ ที่เกิดขึ้น อาจจะเป็นตัวเลข ตัวอักษร หรือสัญลักษณ์ก็ได้ ข้อมูลที่ดีจะต้องมีความถูกต้องแม่นยำ และเป็นปัจจุบัน เช่น ปริมาณ ระยะทาง ชื่อ ที่อยู่ เบอร์โทรศัพท์ คะแนนของนักเรียน รายงาน บันทึก ฯลฯ

เนื่องจากข้อมูลที่เก็บรวบรวมอยู่ในฐานข้อมูล บางส่วนยังเป็นข้อมูลที่ผิดพลาด หรือไม่สมบูรณ์ เช่น ผู้ที่ทำการจัดเก็บข้อมูลใส่ข้อมูลไม่ครบ หรือการจัดเก็บมีความซ้ำซ้อนกัน

ขั้นตอนในการจัดเตรียมข้อมูล

2.2.1 การเลือกข้อมูล (Data selection)

เลือกข้อมูลเฉพาะที่ต้องการ เพื่อที่จะนำมาวิเคราะห์ให้ตรงตามจุดประสงค์ของการจัดกลุ่ม การเลือกข้อมูลจะแตกต่างไปตามวัตถุประสงค์ที่ต้องการของแต่ละธุรกิจ

2.2.2 การเตรียมข้อมูล (Data Preprocessing)

ข้อมูลที่มีอยู่อาจมีบางส่วนที่ผิดพลาดหรือไม่สมบูรณ์เช่น ข้อมูลบางส่วนขาดหายไป ข้อมูลไม่สอดคล้องกัน ข้อมูลมีความซ้ำซ้อน ซึ่งสามารถแก้ไขโดย

- Data Cleaning เป็นการจัดการข้อมูลส่วนที่ขาดหายไปด้วยการใส่ค่า “unknown”
- Data Integration เป็นการกำจัดความซ้ำซ้อนของข้อมูล

2.2.3 การแปลงข้อมูล (Data Transformation)

การแปลงข้อมูลจะเป็นการนำข้อมูลที่มีอยู่มาเปลี่ยนแปลงทำให้อยู่ในรูปแบบของข้อมูลที่พร้อมจะนำไปวิเคราะห์กับอัลกอลิทึม เช่น การแปลงข้อมูลตัวอักษรหรือสัญลักษณ์ให้กลายเป็นตัวเลข เนื่องจากอัลกอลิทึมบางอัลกอลิทึมต้องใช้ข้อมูลที่เป็นตัวเลขเท่านั้น

ตัวอย่าง

ตารางที่ 2.1 ชื่อข้อมูลและชนิดของข้อมูลก่อนแปลง

ชื่อข้อมูล	ชนิดข้อมูล
คณะ	Text
สาขาวิชา	Text

ตารางที่ 2.2 ชนิดข้อมูลของคณะหลังแปลง

ข้อมูลเดิม	ข้อมูลใหม่
เทคโนโลยีสารสนเทศ	1
วิศวกรรมศาสตร์	2
วิทยาศาสตร์	3

ตารางที่ 2.3 ชนิดข้อมูลของสาขาวิชาหลังแปลง

ข้อมูลเดิม	ข้อมูลใหม่
เทคโนโลยีสารสนเทศ	1
เทคโนโลยีสารสนเทศ	2
วิศวกรรมคอมพิวเตอร์	3
วิทยาการคอมพิวเตอร์	4

2.3 การกำหนดความคล้ายคลึงกันในกลุ่มของข้อมูล

การวิเคราะห์การจัดกลุ่มโดยทั่วไปจะใช้ Distance function ซึ่งเป็นฟังก์ชั่นที่ใช้กำหนดความคล้ายคลึงกันในกลุ่มของข้อมูล โดยพิจารณาจากระยะระหว่างข้อมูล 2 ตัวซึ่งจะอยู่ในรูปของ Matrix โดย Matrix แบ่งออกเป็น 2 ชนิด

- กำหนดจากเมทริกซ์ข้อมูล (Data matrix)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

รูปที่ 2.1 เมทริกซ์ข้อมูล

- กำหนดจากเมทริกซ์ความต่าง (Dissimilarity matrix)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

รูปที่ 2.2 เมทริกซ์ความต่าง

ระยะทาง มีความสัมพันธ์ผกผันกับตัววัดความคล้ายคลึง (Similarity มาก ค่าระยะจะน้อย แต่ถ้า Similarity น้อย ค่าระยะจะมาก) การคำนวณที่นิยมใช้ในปัจจุบันคือ Euclidean distance

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn})^2} \quad (2.1)$$

เมื่อ $i = (x_{i1}, x_{i2},\ldots, x_{in})$ และ $j = (x_{j1}, x_{j2},\ldots, x_{jn})$ เป็นข้อมูลขนาด n มิติ

สมบัติของ Matrix

- $d(x, y) \geq 0$ โดย $d(x, y) = 0$ เมื่อ $x = y$ เท่านั้น (ค่าความแตกต่างของข้อมูล ซึ่งเป็นค่าระยะห่างระหว่างจุด 2 จุด จะมีค่าเป็นบวกและจะเป็นศูนย์เมื่อจุด x และ y เป็นจุดเดียวกัน)
- $d(x, y) = d(y, x)$ คือ จะสมมาตรกัน (ค่าความแตกต่างของข้อมูลระหว่าง x และ y จะเหมือนกันในแต่ละทิศทาง)
- $d(x, z) \leq d(x, y) + d(y, z)$ คือ เมื่อพิจารณาค่าความแตกต่างของข้อมูลระหว่างจุด 2 จุดที่สั้นที่สุดจะสั้นกว่าหรือเท่ากับค่าความแตกต่างของข้อมูลของเส้นทางที่ประกอบกันด้วยจำนวนจุดที่มากกว่าโดยจุดต้นและจุดปลายเป็นจุดเดียวกัน

นิยามของค่าความแตกต่างของข้อมูลระหว่างจำนวนจริง x และ y คือ $d(x, y) = |x - y|$ ซึ่งนิยามนี้เป็นไปตามเงื่อนไขทั้ง 3 ข้อด้านบน และตรงกับโครงสร้างมาตรฐานของเส้นจำนวนจริง (Real line)

2.4 อัลกอลิทึมที่ใช้ในการจัดกลุ่มของข้อมูล

อัลกอลิทึมที่ใช้ในการจัดกลุ่มนั้นจะพยายามค้นหาข้อมูลของกลุ่มตามธรรมชาติ ซึ่งหมายความว่าจะต้องหาว่าข้อมูลที่กำลังจะทำการจัดกลุ่มนั้น มีลักษณะเหมือนกันข้อมูลในกลุ่มใดมากที่สุด แล้วก็จะเอาข้อมูลนั้นไปจัดไว้ในกลุ่มที่มีลักษณะเหมือนกันที่สุด ซึ่งเราสามารถแบ่งประเภทของ อัลกอลิทึมที่ใช้ในการจัดกลุ่มได้ดังนี้

- การจัดกลุ่มโดยใช้วิธีแบ่งกลุ่ม (Partitioning Methods)
- การจัดกลุ่มโดยใช้ระดับชั้น (Hierarchical Methods)
- การจัดกลุ่มโดยใช้ความหนาแน่น (Density-Based Methods)
- การจัดกลุ่มโดยใช้วิธีแบ่งเป็นช่องๆ (Grid-Based Methods)
- การจัดกลุ่มโดยใช้ตัวแบบ (Model-Based Methods)

ในโครงงานนี้ขอกล่าวถึงเฉพาะอัลกอลิทึมเป็นที่นิยมมากในปัจจุบัน

2.4.1 การจัดกลุ่มโดยใช้วิธีแบ่งกลุ่ม (Partitioning Methods)

ข้อมูลที่จะนำมาจัดกลุ่มจะถูกนำไปเปรียบเทียบว่าเหมือนกับกลุ่มใดมากที่สุด แล้วจึงนำข้อมูลนั้นไปจัดไว้ในกลุ่มที่มีลักษณะเหมือนกันที่สุด โดยอัลกอริทึมที่ใช้มีด้วยกันหลายอัลกอริทึม

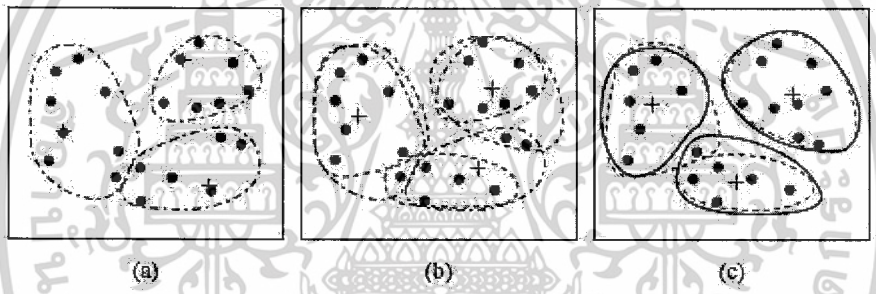
2.4.1.1 K-means algorithm

จัดกลุ่มโดยการกำหนดจำนวนกลุ่มที่ต้องการใช้ในการจัดกลุ่มล่วงหน้า ซึ่งจำนวนกลุ่มที่ใช้ในการจัดกลุ่มนี้จะแทนด้วยตัวแปร k อัลกอริทึมนี้เป็นการสร้างกลุ่มของข้อมูลโดยค่าเฉลี่ยของข้อมูลในกลุ่มจะถูกกำหนดให้เป็นศูนย์กลางของกลุ่ม โดยศูนย์กลางของแต่ละกลุ่มจะแสดงโดยค่าเฉลี่ยของข้อมูลในกลุ่มนั้น

Input: k = จำนวนกลุ่มที่ต้องการจะแบ่ง
n = จำนวนข้อมูลที่ใช้ในการแบ่งกลุ่ม
Output: ข้อมูลที่ถูกแบ่งออกเป็น k กลุ่ม

วิธีและหลักการทำงาน

1. เลือกข้อมูล k ตัวจากข้อมูลทั้งหมดเพื่อเป็นศูนย์กลางของกลุ่ม ซึ่ง k คือจำนวนของกลุ่มที่จะแบ่ง (รูป a)
2. จัดข้อมูลที่เหลือเข้ากลุ่มข้อมูล โดยเปรียบเทียบค่าของข้อมูลใกล้เคียงกับศูนย์กลางของกลุ่มนั้นมากที่สุด (รูป a)
3. ทำการคำนวนศูนย์กลางของกลุ่มที่มีการเปลี่ยนแปลงข้อมูลใหม่ แล้วคำนวณเปรียบเทียบค่าของข้อมูลแต่ละตัวอีกครั้ง โดยเปรียบเทียบกับศูนย์กลางของแต่ละกลุ่ม ถ้าข้อมูลตัวใดมีค่าของข้อมูลใกล้เคียงกับศูนย์กลางของกลุ่มอื่นมากกว่า ให้ย้ายข้อมูลไปอยู่ในกลุ่มนั้น (รูป b)
4. ทำซ้ำข้อ 3 จนไม่มีการเปลี่ยนแปลงข้อมูลของกลุ่ม หรือไม่มีการเปลี่ยนแปลงศูนย์กลางของกลุ่ม (รูป c)



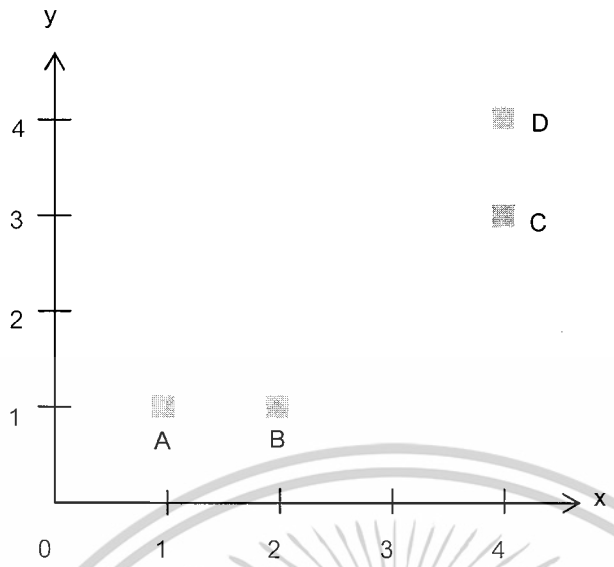
รูปที่ 2.3 การจัดกลุ่มของข้อมูลโดยใช้ K-means algorithm

ตัวอย่างการจัดกลุ่มข้อมูลโดยใช้วิธี K-means algorithm

สมมติให้มี Object อยู่ 4 Object ซึ่งแต่ละ Object มี 2 Attribute ดังแสดงในตารางที่ 3.1 พิกัดของ Object ต่างๆแสดง ดังรูปที่ 3.1 และต้องการแบ่งกลุ่มออกเป็น 2 กลุ่ม

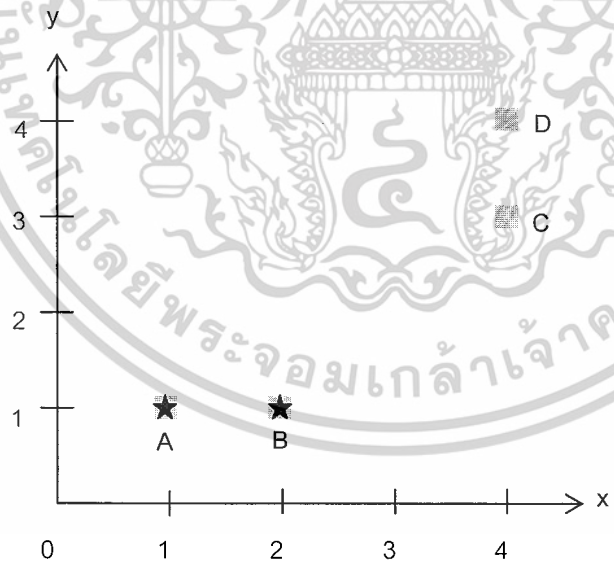
ตารางที่ 2.4 ตารางข้อมูลของ Object

Object	Attribute1 (x)	Attribute2 (y)
A	1	1
B	2	1
C	4	3
D	4	4



รูปที่ 2.4 พิกัดของ Object ต่างๆ

ขั้นที่ 1 เลือกข้อมูล 2 ตัวจากข้อมูลทั้งหมดเพื่อนำมาเป็นตัวแทนของกลุ่ม ในที่นี้กำหนดให้ A และ B เป็นตัวแทนของกลุ่มที่ 1 และ 2 ตามลำดับ โดยศูนย์กลางของกลุ่มที่ 1 คือ (1 ,1) และศูนย์กลางของกลุ่มที่ 2 คือ (2 ,1) ดังรูป 3.2



รูปที่ 2.5 พิกัดของ Object ต่างๆ และจุดศูนย์กลางเริ่มต้นของแต่ละกลุ่ม

ขั้นที่ 2 คำนวณหาค่า Distance หรือระยะห่างระหว่าง Object แต่ละตัวนำไปเปรียบเทียบกับศูนย์กลางของแต่ละกลุ่ม แล้วทำการจัดข้อมูลเข้ากลุ่มโดยจัด Object เข้ากลุ่มที่มีค่า Distance น้อยที่สุด

Object C เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(3-1)^2} = 3.61$

Object C เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-2)^2+(3-1)^2} = 2.83$

Object D เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(4-1)^2} = 4.24$

Object D เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-2)^2+(4-1)^2} = 3.61$

ตารางที่ 2.5 ตารางข้อมูลของ Object

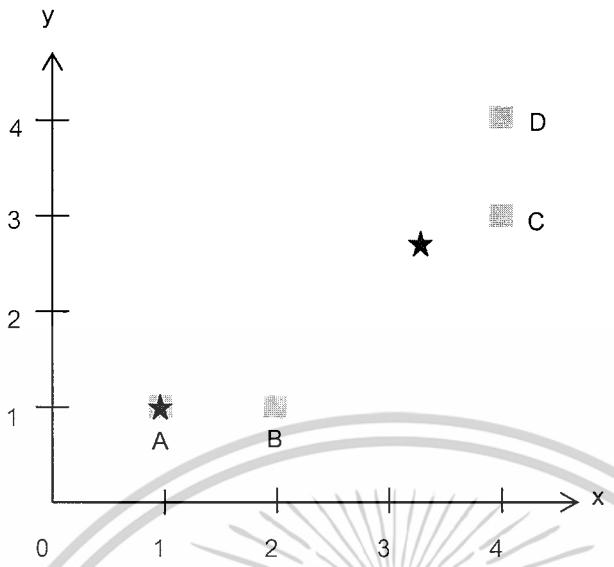
Object	Attribute1 (x)	Attribute2 (y)	Distance Group 1	Distance Group 2	Group
A	1	1	0	1	1
B	2	1	1	0	2
C	4	3	3.61	2.83	2
D	4	4	4.24	3.61	2

ขั้นที่ 3 คำนวณหาศูนย์กลางของกลุ่มใหม่อีกครั้ง โดยกลุ่มที่ 1 ไม่มีการเปลี่ยนแปลงของสมาชิกในกลุ่ม ดังนั้นจุดศูนย์กลางของกลุ่มยังคงเป็น (1 ,1) ส่วนในกลุ่มที่ 2 ซึ่งมีสมาชิกเพิ่มเข้ามาเป็น 3 Object ทำให้ต้องคำนวณหาศูนย์กลางของกลุ่มใหม่ได้พิกัดดังรูปที่ 3.2 แล้วทำซ้ำขั้นที่ 2 อีกครั้ง ได้ผลดังตารางที่ 3.2

Center Group 1 = (1,1)

$$\text{Center Group 2} = \left(\frac{2+4+4}{3}, \frac{1+3+4}{3} \right)$$

$$= \left(\frac{10}{3}, \frac{8}{3} \right)$$



รูปที่ 2.6 พิกัดของ Object ต่างๆ และจุดศูนย์กลางหลังการเปลี่ยนครั้งที่ 1

ตารางที่ 2.6 ตารางข้อมูลของ Object หลังการเปลี่ยนครั้งที่ 1

Object	Attribute1 (x)	Attribute2 (y)	Distance Group 1	Distance Group 2	Group
A	1	1	0	3.36	1
B	2	1	1	2.75	1
C	4	3	3.61	0.74	2
D	4	4	4.24	1.48	2

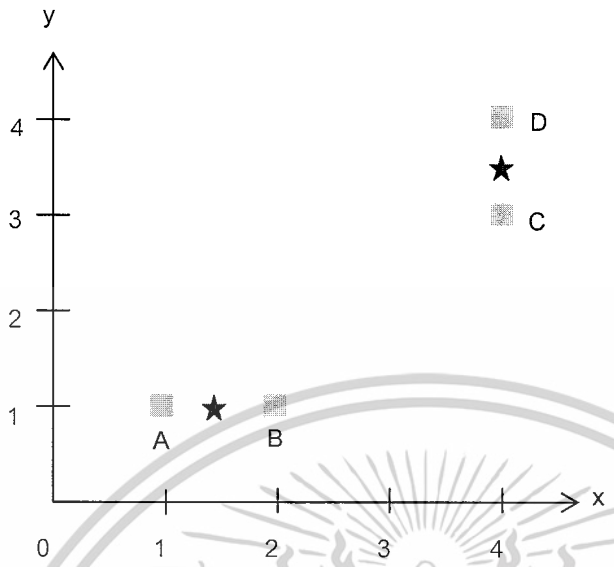
ขั้นที่ 4 ทำซ้ำขั้นตอนที่ 3 ได้จุดศูนย์กลางใหม่ดังรูปที่ 3.3 แล้วทำซ้ำขั้นที่ 2 ได้ผลดังตารางที่ 3.3

$$\text{Center Group 1} = \left(\frac{1+2}{2}, \frac{1+1}{2}\right)$$

$$= \left(\frac{10}{2}, \frac{8}{2}\right)$$

$$\text{Center Group 2} = \left(\frac{4+4}{2}, \frac{3+4}{2}\right)$$

$$= \left(4, \frac{7}{2}\right)$$



รูปที่ 2.7 พิกัดของ Object ต่างๆ และจุดศูนย์กลางหลังการเปลี่ยนครั้งที่ 2

ตารางที่ 2.7 ตารางข้อมูลของ Object หลังการเปลี่ยนครั้งที่ 2

Object	Attribute1 (x)	Attribute2 (y)	Distance Group 1	Distance Group 2	Group
A	1	1	0.50	3.95	1
B	2	1	0.50	3.20	1
C	4	3	3.20	0.50	2
D	4	4	3.20	0.50	2

กลุ่มของ Object หลังการเปลี่ยนครั้งที่ 2 เหมือนกันกับการเปลี่ยนครั้งที่ 1 แสดงว่าไม่มีการเปลี่ยนกลุ่มของ Object เกิดขึ้น ดังนั้นจึงนำเอากลุ่มที่ได้จากการเปลี่ยนกลุ่มครั้งสุดท้ายของ Object มาเป็นผลลัพธ์ของการจัดกลุ่ม ซึ่งก็คือตารางที่ 3.3

2.4.1.2 K-medoids algorithm

จัดกลุ่มโดยการกำหนดจำนวนกลุ่มที่ต้องการใช้ในการจัดกลุ่มล่วงหน้า ซึ่งจำนวนกลุ่มที่ใช้ในการจัดกลุ่มนี้จะแทนด้วยตัวแปร k อัลกอริทึมนี้เป็นการสร้างกลุ่มของข้อมูลโดยกำหนดให้ข้อมูลตัวใดตัวหนึ่งเป็นตัวแทนของกลุ่ม

2.4.1.2.1 PAM (Partitioning Around Medoids)

Input: k = จำนวนกลุ่มที่ต้องการจะแบ่ง

n = จำนวนข้อมูลที่ใช้ในการแบ่งกลุ่ม

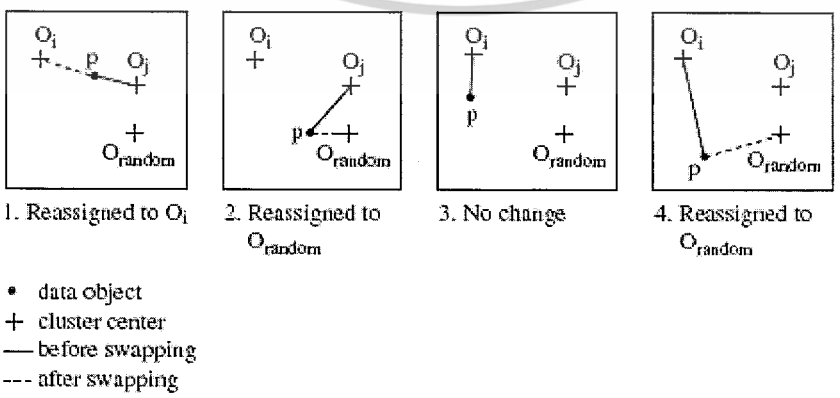
D = ข้อมูลทั้งหมดที่ต้องการแบ่งกลุ่ม

Output: ข้อมูลที่ถูกแบ่งออกเป็น k กลุ่ม

วิธีและหลักการทำงาน

1. เลือกข้อมูล k ตัวจากข้อมูลทั้งหมดเพื่อเป็นตัวแทนของกลุ่ม ซึ่ง k คือจำนวนของกลุ่มที่จะแบ่ง
2. จัดข้อมูลที่เหลือเข้ากลุ่มข้อมูล โดยเปรียบเทียบค่าของข้อมูลใกล้เคียงกับศูนย์กลางของกลุ่มนั้นมากที่สุด
3. เลือกข้อมูลตัวอื่นที่ไม่ใช่ตัวแทนโดยการสุ่มเพื่อนำมาเปรียบเทียบตัวแทนโดยคำนวณค่าของการสลับ(S) ถ้า S<0 ให้ทำการเปลี่ยนข้อมูลที่สุ่มมาเป็นตัวแทน
4. ทำซ้ำข้อ 3 จนไม่มีการเปลี่ยนตัวแทน

โดยค่าของการสลับ(S) คำนวณจาก ผลรวมของผลต่างค่า Distance จากกรณีต่างๆต่อไปนี้ ซึ่งเมื่อสุ่มตัวแทนจากข้อมูลตัวอื่นที่ไม่ใช่ตัวแทน เพื่อนำมาแทนตัวแทนในกลุ่มที่สุ่มขึ้นมา แล้วทำการคำนวณหา Distance ใหม่โดยสามารถแยกเป็น 4 กรณีดังรูปที่ 2.4



รูปที่ 2.8 การจัดกลุ่มของข้อมูลโดยใช้ K-medoid algorithm

กรณีที่ 1 ถ้าระยะห่างระหว่าง Object ที่อยู่ในกลุ่มเดียวกับตัวแทนใหม่ที่ถูกสุ่มขึ้นมาแทนที่มีค่ามากกว่า ระยะห่างระหว่าง Object นั้นไปตัวแทนในกลุ่มอื่นๆ ให้ย้าย Object นั้นไปยังกลุ่มที่มีค่าน้อยที่สุด แล้วนำค่า Distance ใหม่ที่มีการเปลี่ยนตัวแทนแล้ว ลบกับค่า Distance เก่าตอนที่ยังไม่ได้เปลี่ยนตัวแทน ซึ่งในกรณีนี้จะได้ค่าเป็นบวก

กรณีที่ 2 ถ้าระยะห่างระหว่าง Object ที่อยู่ในกลุ่มเดียวกับตัวแทนใหม่ที่ถูกสุ่มขึ้นมาแทนที่มีค่าน้อยกว่า ระยะห่างระหว่าง Object นั้นไปยังตัวแทนเดิมในกลุ่ม จะให้ Object นั้นอยู่ในกลุ่มเดิมโดยเปลี่ยนไปชี้ที่ตัวแทนใหม่ แล้วนำค่า Distance ใหม่ที่มีการเปลี่ยนตัวแทนแล้ว ลบกับค่า Distance เก่าตอนที่ยังไม่ได้เปลี่ยนตัวแทน ซึ่งในกรณีนี้จะได้ค่าเป็นลบ

กรณีที่ 3 ถ้าระยะห่างระหว่าง Object ที่อยู่ต่างกลุ่มกับตัวแทนใหม่ที่ถูกสุ่มขึ้นมาแทนที่มีค่ามากกว่า ระยะห่างระหว่าง Object นั้นไปยังตัวแทนในกลุ่มที่ Object นั้นอยู่ จะไม่มีการเปลี่ยนแปลงเกิดขึ้น ดังนั้นกรณีนี้จะได้ค่าเป็น 0 เนื่องจากไม่มีการเปลี่ยนแปลง

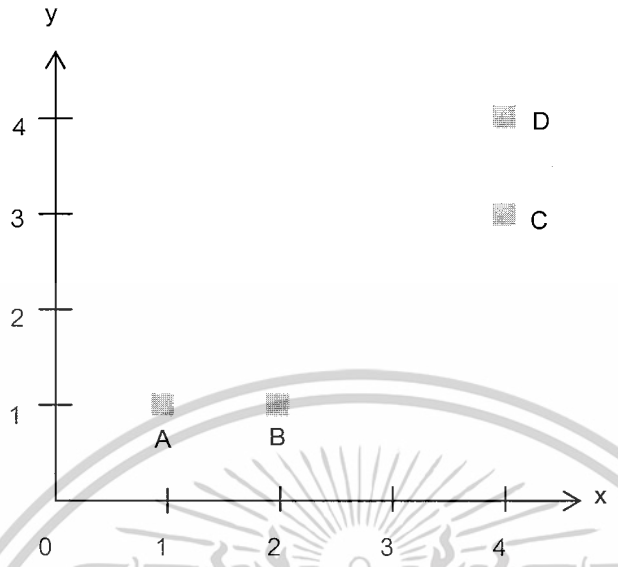
กรณีที่ 4 ถ้าระยะห่างระหว่าง Object ที่อยู่ต่างกลุ่มกับตัวแทนใหม่ที่ถูกสุ่มขึ้นมาแทนที่มีค่าน้อยกว่า ระยะห่างระหว่าง Object นั้นไปยังตัวแทนในกลุ่มที่ Object นั้นอยู่ ให้ย้าย Object นั้นไปยังกลุ่มที่มีค่าน้อยที่สุด แล้วนำค่า Distance ใหม่ที่มีการเปลี่ยนตัวแทนแล้ว ลบกับค่า Distance เก่าตอนที่ยังไม่ได้เปลี่ยนตัวแทน ซึ่งในกรณีนี้จะได้ค่าเป็นลบ

ตัวอย่างการจัดกลุ่มข้อมูลโดยใช้วิธี K-medoids algorithm (PAM)

สมมติให้มี Object อยู่ 4 Object ซึ่งแต่ละ Object มี 2 Attribute ดังแสดงในตารางที่ 2.8 พิกัดของ Object ต่างๆแสดง ดังรูปที่ 2.9 และต้องการแบ่งกลุ่มของ Object ออกเป็น 2 กลุ่ม โดยใช้ K-means algorithm

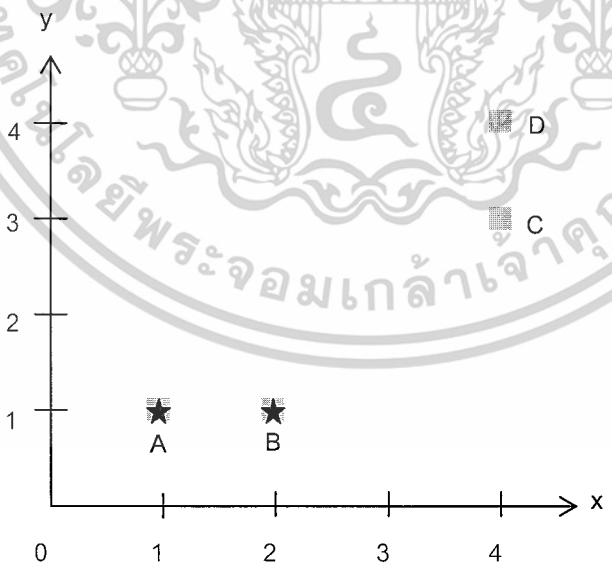
ตารางที่ 2.8 ตารางข้อมูลของ Object

Object	Attribute1 (x)	Attribute2 (y)
A	1	1
B	2	1
C	4	3
D	4	4



รูปที่ 2.9 พิกัดของ Object ต่างๆ

ขั้นที่ 1 เลือกข้อมูล 2 ตัวจากข้อมูลทั้งหมดเพื่อนำมาเป็นตัวแทนของกลุ่ม ในที่นี้กำหนดให้ A และ B เป็นตัวแทนของกลุ่มที่ 1 และ 2 ตามลำดับ โดยตัวแทนของกลุ่มที่ 1 คือ (1 ,1) และตัวแทนของกลุ่มที่ 2 คือ (2 ,1) ดังรูป 2.10



รูปที่ 2.10 พิกัดของ Object ต่างๆ และตัวแทนของแต่ละกลุ่ม

ขั้นที่ 2 คำนวณหาค่า Distance หรือระยะห่างระหว่าง Object แต่ละตัวนำไปเปรียบเทียบกับศูนย์กลางของแต่ละกลุ่ม แล้วทำการจัดข้อมูลเข้ากลุ่มโดยจัด Object เข้ากลุ่มที่มีค่า Distance น้อยที่สุด

Object C เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(3-1)^2} = 3.61$

Object C เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-2)^2+(3-1)^2} = 2.83$

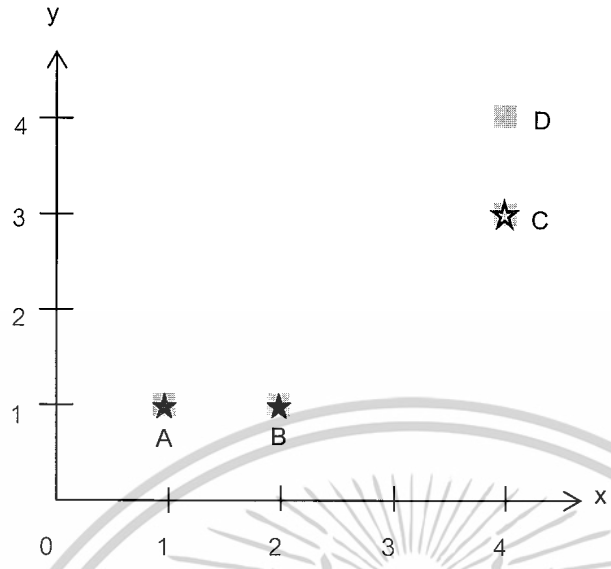
Object D เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(4-1)^2} = 4.24$

Object D เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-2)^2+(4-1)^2} = 3.61$

ตารางที่ 2.9 ตารางข้อมูลของ Object หลังการจัดกลุ่ม

Object	Attribute1 (x)	Attribute2 (y)	Distance Group 1	Distance Group 2	Group
A	1	1	0	1	1
B	2	1	1	0	2
C	4	3	3.61	2.83	2
D	4	4	4.24	3.61	2

ขั้นที่ 3 ทำการสุ่ม Object ที่ไม่ใช่ตัวแทนเพื่อ เอามาเป็นตัวแทนใหม่ ในที่นี้กำหนดให้ C คือ Object ที่ถูกสุ่มขึ้นมาใหม่ ดังรูปที่ 2.11 ดังนั้น C จึงถูกกำหนดให้เป็นตัวแทนใหม่ของกลุ่มที่ 2 คำนวณหาค่า Distance ใหม่อีกครั้ง ได้ดังตารางที่ 2.10



รูปที่ 2.11 พิกัดของ Object ต่างๆ ตัวแทนของแต่ละกลุ่ม และตัวแทนใหม่

Object A เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(1-1)^2+(1-1)^2} = 0$

Object A เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(1-4)^2+(1-3)^2} = 3.61$

Object B เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(2-1)^2+(1-1)^2} = 1$

Object B เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(2-4)^2+(1-3)^2} = 2.83$

Object C เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(3-1)^2} = 3.61$

Object C เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-4)^2+(3-3)^2} = 0$

Object D เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(4-1)^2} = 4.24$

Object D เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-4)^2+(4-3)^2} = 1$

ตารางที่ 2.10 ตารางข้อมูลของ Object หลังการเปลี่ยนตัวแทนครั้งที่ 1

Object	Attribute1 (x)	Attribute2 (y)	Distance Group 1	Distance Group 2	Group
A	1	1	0	3.61	1
B	2	1	1	2.83	1
C	4	3	3.61	0	2
D	4	4	4.24	1	2

คำนวณ ค่าของการสลับ(S) โดยแยกเป็น 4 กรณีดังที่กล่าวไปแล้ว

จากตัวอย่าง Object A เข้ากรณีที่ 3 ดังนั้น ค่าของการสลับ(S) = 0

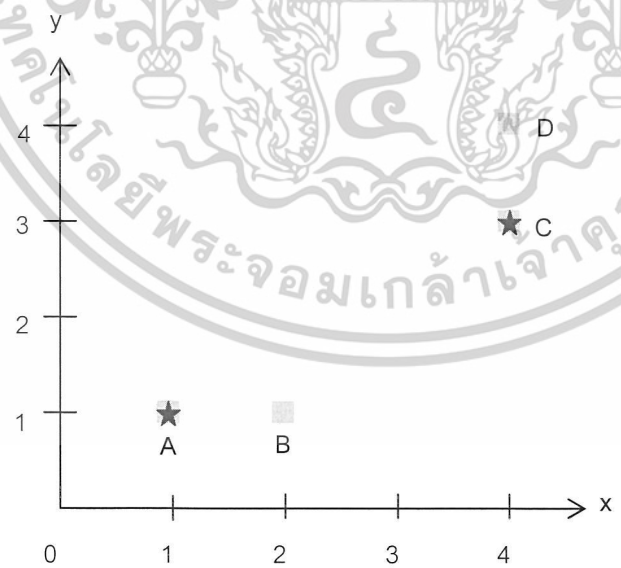
Object B เข้ากรณีที่ 1 ดังนั้น ค่าของการสลับ(S) = 1 – 0 = 1

Object C เข้ากรณีที่ 2 ดังนั้น ค่าของการสลับ(S) = 0 – 2.83 = - 2.83

Object D เข้ากรณีที่ 2 ดังนั้น ค่าของการสลับ(S) = 1 – 3.61 = - 2.61

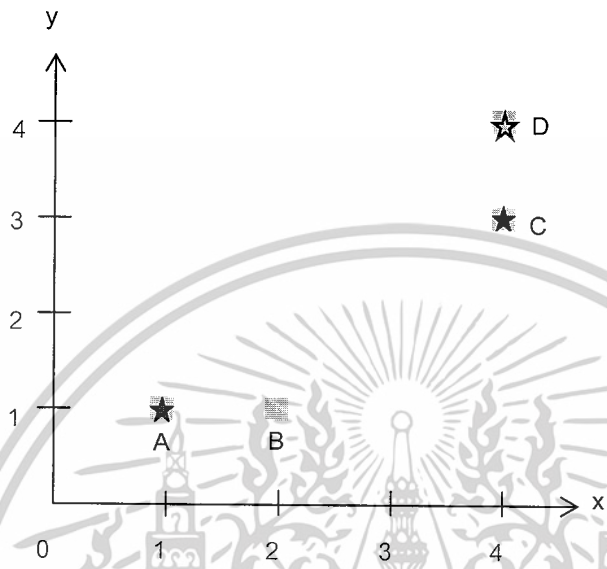
ผลรวมของ ค่าของการสลับ(S) = 0 + 1 + (- 2.83) + (- 2.61) = - 4.44

ซึ่ง – 4.44 < 0 ทำการสลับตัวแทนของกลุ่ม 2 ดังรูป 2.12



รูปที่ 2.12 พิกัดของ Object ต่างๆ และตัวแทนของแต่ละกลุ่ม

ขั้นที่ 4 ทำซ้ำขั้นที่ 3 ทำการสุ่ม Object ที่ไม่ใช่ตัวแทนเพื่อ เอามาเป็นตัวแทนใหม่ ในที่นี้ กำหนดให้ D คือ Object ที่ถูกสุ่มขึ้นมาใหม่ ดังรูป 2.13 ดังนั้น D จึงถูกกำหนดให้เป็นตัวแทนใหม่ ของกลุ่มที่ 2 คำนวณหาค่า Distance ใหม่อีกครั้ง ได้ดังตารางที่ 2.11



รูปที่ 2.13 พิกัดของ Object ต่างๆ ตัวแทนของแต่ละกลุ่ม และตัวแทนใหม่

Object A เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(1-1)^2+(1-1)^2} = 0$

Object A เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(1-4)^2+(1-4)^2} = 4.24$

Object B เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(2-1)^2+(1-1)^2} = 1$

Object B เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(2-4)^2+(1-4)^2} = 3.61$

Object C เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(3-1)^2} = 3.61$

Object C เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-4)^2+(3-4)^2} = 1$

Object D เปรียบเทียบกับ กลุ่มที่ 1 $\sqrt{(4-1)^2+(4-1)^2} = 4.24$

Object D เปรียบเทียบกับ กลุ่มที่ 2 $\sqrt{(4-4)^2+(4-4)^2} = 0$

ตารางที่ 2.11 ตารางข้อมูลของ Object หลังการเปลี่ยนตัวแทนครั้งที่ 2

Object	Attribute1 (x)	Attribute2 (y)	Distance Group 1	Distance Group 2	Group
A	1	1	0	4.24	1
B	2	1	1	3.61	1
C	4	3	3.61	1	2
D	4	4	4.24	0	2

คำนวณ ค่าของการสลับ(S) โดยแยกเป็น 4 กรณีดังที่กล่าวไปแล้ว

จากตัวอย่าง Object A เข้ากรณีที่ 3 ดังนั้น ค่าของการสลับ(S) = 0

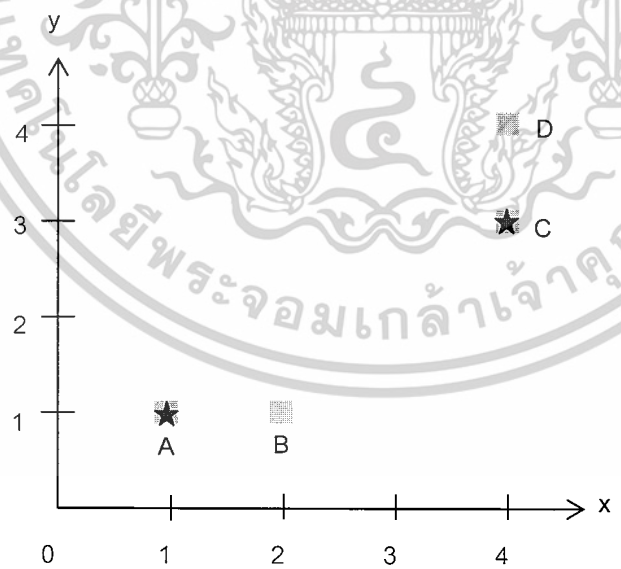
Object B เข้ากรณีที่ 3 ดังนั้น ค่าของการสลับ(S) = 0

Object C เข้ากรณีที่ 2 ดังนั้น ค่าของการสลับ(S) = 1 – 0 = 1

Object D เข้ากรณีที่ 2 ดังนั้น ค่าของการสลับ(S) = 0 – 1 = - 1

ผลรวมของ ค่าของการสลับ(S) = 0 + 0 + 1 + (- 1) = 0

ดังนั้นจึงไม่มีการเปลี่ยนตัวแทน และ Object ทุกตัวถูกสุ่มมาหมดแล้ว ผลที่ได้จากการคำนวณเป็นดังรูปที่ 2.14 และตารางที่ 2.12



รูปที่ 2.14 พิกัดของ Object ต่างๆ และตัวแทนของแต่ละกลุ่ม

ตารางที่ 2.12 ตารางข้อมูลของ Object หลังการคำนวณ

Object	Attribute1 (x)	Attribute2 (y)	Distance Group 1	Distance Group 2	Group
A	1	1	0	3.61	1
B	2	1	1	2.83	1
C	4	3	3.61	0	2
D	4	4	4.24	1	2

2.1.4.1.2 CLARA algorithm

ถูกพัฒนาขึ้นมาจาก PAM algorithm เพื่อแก้ปัญหาการประมวณผลที่นานของการทำงานของ PAM algorithm โดยการเลือกตัวเปลี่ยนกับตัวแทน จะทำโดยกำหนดปริมาณของข้อมูลที่จะถูกสุ่มขึ้นมาเป็นตัวแทน เป็น 40+2k

วิธีและหลักการทำงาน

1. เลือกข้อมูล 40+2k ตัวจากข้อมูลทั้งหมดโดยการสุ่ม
2. เลือกข้อมูล k ตัวจากข้อมูลที่สุ่มขึ้นมาเป็นตัวแทนของกลุ่ม ซึ่ง k คือจำนวนของกลุ่มที่จะแบ่ง
3. ทำการเลือก k ที่เป็นตัวแทนที่ดีที่สุดโดยใช้อัลกอริทึม PAM
4. นำ k ที่ได้จากขั้นตอนที่ 3 คำนวณหาประสิทธิภาพในการจัดกลุ่มของข้อมูลทั้งหมด
5. ทำซ้ำข้อ 1 วนทำจนครบ 5 ครั้ง โดยในขั้นตอนที่ 4 ให้คำนวณหาประสิทธิภาพในการจัดกลุ่มมาเปรียบเทียบกับครั้งก่อนหน้า ถ้าการจัดกลุ่มครั้งหลังมีประสิทธิภาพมากกว่าให้นำ k ที่ได้จากครั้งหลังเป็นตัวหลักต่อไป

2.1.4.1.3 CLARANS algorithm

ถูกพัฒนาขึ้นมาจาก PAM algorithm เพื่อแก้ปัญหาการประมวณผลที่นานของการทำงานของ PAM algorithm โดยจะทำการสุ่มตัวแทนของกลุ่มขึ้นมา k ตัวจากข้อมูลทั้งหมดแล้วทำการสลับเปลี่ยนตัวแทนที่สุ่มขึ้นมาทีละตัวโดยสุ่มจากข้อมูลทั้งหมดใหม่ เพื่อหาตัวแทนที่ดีที่สุด

วิธีและหลักการทำงาน

1. เลือกข้อมูล k ตัวจากข้อมูลทั้งหมดโดยการสุ่ม เพื่อเป็นตัวแทนของกลุ่ม
2. คำนวณหาประสิทธิภาพในการจัดกลุ่ม
3. ทำการสุ่มเลือกตัวแทน 1 ตัวจากข้อมูลทั้งหมด โดยนำมาเปลี่ยนกับตัวแทนที่สุ่มได้ในขั้นตอนที่ 1 แล้วทำการคำนวณหาประสิทธิภาพในการจัดกลุ่มใหม่แล้วเปรียบเทียบประสิทธิภาพในการจัดกลุ่ม นำตัวแทนชุดที่มีประสิทธิภาพในการจัดกลุ่มดีกว่าเป็นตัวหลักต่อไป
4. ทำซ้ำข้อ 3 จนครบจำนวณรอบที่กำหนดไว้ในการเปลี่ยนตัวแทนกลุ่ม
5. ทำซ้ำข้อ 1 จนครบจำนวณรอบที่กำหนดไว้ในการสุ่มตัวแทนกลุ่มใหม่ ถ้าการสุ่มใหม่ได้ประสิทธิภาพของการจัดกลุ่มดีกว่า ให้นำตัวแทนที่ได้จากการสุ่มใหม่เป็นตัวหลัก

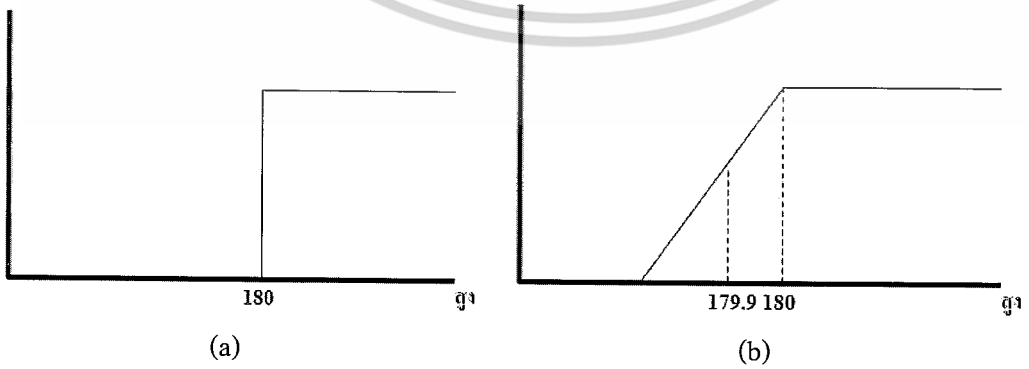
2.4.1.4 Fuzzy Set

มนุษย์มีความสามารถในการจดจำ วิเคราะห์ แยกแยะ และจัดแบ่งข้อมูล แต่บ่อยครั้งที่การทำงานเหล่านี้มีข้อจำกัดเนื่องจากประสาทสัมผัสของมนุษย์ไม่สามารถรับรู้แยกแยะข้อมูลบางอย่างได้ทั้งหมด ทำให้ไม่สามารถตัดสินใจได้ว่า ข้อมูลที่ได้รับเข้ามานั้นมีรูปแบบและลักษณะที่ชัดเจนอย่างไร ทำให้เกิดความคลุมเครือ ดังนั้นการวิเคราะห์จัดการแบ่งข้อมูลเพื่อให้มีลักษณะที่ชัดเจน จึงมีส่วนสำคัญในกระบวนการวิเคราะห์และจัดการแบ่งข้อมูล

นิยามของฟัซซี่เซต โดยทั่วไป เซตที่เรากล่าวถึง เรามักนิยามเซตที่มีขอบเขตของสมาชิกที่แน่นอน คือเมื่อเราพิจารณาแล้วสามารถเข้าใจว่าข้อมูลตัวใดเป็นสมาชิกหรือไม่เป็นสมาชิกของเซตนั้นบ้าง เช่น กำหนดให้ Set A เป็นเซตของจำนวนจริงที่มีค่ามากกกว่า 5 ดังสมการ

$$A = \{x \mid x > 5\} \qquad (2.2)$$

จากสมการ ข้อมูลที่มีค่ามากกว่า 5 จะเป็นสมาชิกของ Set A เช่น ค่า “6” แต่ถ้าข้อมูลที่มีค่าน้อยกว่า 5 จะไม่เป็นสมาชิกของ Set A เช่น ค่า “4” แต่การนิยามเซตในลักษณะนี้ บางครั้งก็ไม่สามารถอธิบายการทำงานบางอย่างได้ เช่น กำหนดให้ Set A เป็นเซตของคนสูง โดยสมาชิกใน Set A มีค่าเป็นส่วนสูงของแต่ละคน จากนิยามเซตแบบนี้ ทำให้เกิดความคลุมเครือในการบ่งบอกสมาชิกในเซต เช่น ถ้าคนที่มีส่วนสูงมากกว่า 180 cm ถือว่าเป็นคนสูง คนที่มีส่วนสูงน้อยกว่า 180 cm ถือว่าเป็นคนเตี้ย แต่เมื่อพิจารณาคนที่ถูกจัดอยู่ในกลุ่มเตี้ยที่สูง 179.99 cm ซึ่งดูคลุมเครือมาก เพราะมีส่วนสูงต่างกับคนที่สูง 180 cm เพียง 0.01 cm จากปัญหาข้างต้น จึงได้นิยามฟัซซี่เซต (Fuzzy Set) ซึ่งเป็นเซตอีกรูปแบบหนึ่งเพื่อใช้ในการแก้ปัญหาที่พบเหล่านี้ โดยจะมีการกำหนดลักษณะความเป็นสมาชิกให้กับข้อมูลแต่ละตัว ซึ่งลักษณะความเป็นสมาชิกนี้เรียกว่า Membership Function



รูปที่ 2.15 (a) ฟังก์ชั่นของคนสูง (b) ฟังก์ชั่นของฟัซซี่เซตของคนสูง

2.4.1.4.1 ความเป็นสมาชิก (Membership)

ลักษณะความเป็นสมาชิกหรือ Membership Function เป็นค่าที่ใช้บ่งบอกถึงระดับความเป็นสมาชิกของข้อมูลแต่ละตัว ดังนั้น ข้อมูลแต่ละตัวก็จะมีค่าลักษณะความเป็นสมาชิกที่แตกต่างกันไป สำหรับค่าที่ใช้ในการบอกระดับความเป็นสมาชิกจะใช้ค่าความน่าจะเป็นในการเป็นสมาชิกของข้อมูล โดยเรียกค่านี้ว่า Membership Grade โดยค่าความน่าจะเป็นนี้จะอยู่ในช่วง [0 , 1]

2.4.1.4.2 ฟัซซี่ซีมีน (Fuzzy C means Algorithm)

เป็นเทคนิคหนึ่งซึ่งนิยมใช้กันอย่างแพร่หลาย สำหรับการจำแนกข้อมูลที่มีความคลุมเครือออกเป็นกลุ่มๆ เทคนิคนี้ถูกพัฒนาและปรับปรุงมาจาก Dunn's Algorithm โดยศาสตราจารย์ Jim Bezdek โดยลักษณะพิเศษของ Fuzzy C means Algorithm ที่ต่างจากวิธีการอื่นๆ คือ สามารถบ่งบอกถึงระดับความเป็นสมาชิก (Membership Grade) ของแต่ละกลุ่ม (Clusters) ได้

วิธีการวัดและประเมินผลเพื่อให้ทราบว่าข้อมูลแต่ละตัวจัดอยู่ใน Cluster กลุ่มใด จะคำนวณโดยลดค่า Cost Function ผลลัพท์ที่ได้จะอยู่ในรูปของ Matrix อธิบายได้ดังนี้

กำหนดให้

$$X = \{x_1 , x_2 , x_3 , \dots , x_n\} \qquad (2.3)$$

$$V = \{v_1 , v_2 , v_3 , \dots , v_c\} \qquad (2.4)$$

$$J(U,V) = \sum_{j=1}^{C}\sum_{i=1}^{N}(u_{ij})^m \| x_i - v_j \|^2 \qquad (2.5)$$

โดยที่ - X คือ กลุ่มของข้อมูลเข้า ซึ่งมีลักษณะเป็น Input Vectors ที่มีขนาด n ตัว

- c คือจำนวนกลุ่มที่ต้องการแบ่ง
- v แทน Cluster ที่ทำการแบ่ง
- U เป็น matrix ประกอบไปด้วยสมาชิก u_{ij} ซึ่งเป็นค่าบอกระดับความเป็นสมาชิกของข้อมูลตัวที่ i กับ Cluster ที่ j สำหรับ matrix U จะมีมิติเป็น NxC มิติ
- ค่าของสมาชิกที่อยู่ใน Matrix U จะต้องอยู่ในช่วง [0 ,1] เสมอ
- ผลรวมของระดับความเป็นสมาชิก (Membership Grade) ของข้อมูลตัวที่ i ในทุก Cluster มีค่าเท่ากับ 1 เสมอ
- m เป็นตัวแปรที่ใช้กำหนดระดับของความเป็น Fuzzy ซึ่งมีค่าอยู่ในช่วง [0 , ∞)
- $\| x_i - v_j \|$ คือการคำนวณระยะห่าง หรือระยะทางจากข้อมูลตัวที่ i ไปยัง Cluster ที่ j ซึ่งเป็นการคำนวณแบบ Euclidean Distance

วิธีและหลักการทำงาน

1. ทำการสุ่มข้อมูล k ตัวเพื่อเป็นตัวแทนของกลุ่ม กำหนดค่าระดับการเป็น Fuzzy หรือค่าตัวแปร m ที่เหมาะสม และจำนวนรอบ
2. ทำการคำนวณหาค่า Distance ระหว่างข้อมูลกับจุดศูนย์กลางของแต่ละ Cluster โดยใช้สูตร

$$d_{ij} = \| x_i - v_j \| = \sqrt{(x - v_1)^2 + (y - v_2)^2} \qquad (2.6)$$

3. คำนวณหาความเป็นสมาชิกของข้อมูลต่างๆ โดยใช้สูตร

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}} \qquad (2.9)$$

โดย d_{ij} คือ ระยะจาก x ถึง Cendroid v_i ซึ่งกำลังพิจารณาความเป็นสมาชิก

d_{ik} คือ ระยะจาก x ถึง Cendroid v_j ซึ่งเป็น Cendroid ของ Cluster อื่นๆ

4. ทำการคำนวณหาค่าของจุดศูนย์กลางของแต่ละ Cluster ใหม่โดยใช้สูตร

$$v_j = \frac{\sum_{i=1}^{N} (u_{ij})^m x_i}{\sum_{i=1}^{N} (u_{ij})^m} \qquad (2.6)$$

$$J = \{1 , 2 , 3 , \dots , c\} \qquad (2.7)$$

5. กลับไปทำข้อ 2 ใหม่ จนกว่าจุดศูนย์กลางของแต่ละ Cluster ไม่เปลี่ยนแปลง หรือครบตามจำนวนรอบ
6. ตรวจสอบข้อมูลแต่ละตัวว่ามีค่าของความเป็นสมาชิกหรือ Membership Function ของ Cluster กลุ่มใดสูงที่สุด แสดงว่าข้อมูลตัวนั้นน่าจะอยู่ในกลุ่ม Cluster นั้น

2.4.1.5 Self Organizing Map (SOM)

Self Organizing Map (SOM) ถูกพัฒนาโดยศาสตราจารย์โคโฮเนน ซึ่งเป็นอัลกอริทึมที่ได้รับความนิยมอย่างแพร่หลาย และได้มีการนำไปประยุกต์ใช้อย่างแพร่หลาย

2.4.1.5.1 ประเภทของนิวรอนเน็ตเวิร์ค

กระบวนการทำงานของนิวรอนเน็ตเวิร์คแบ่งออกเป็น 3 ประเภทหลักๆ ได้ดังนี้

1. นิวรอนแบบส่งค่าต่อ (Feed Forward Network) เป็นการส่งกลุ่มข้อมูลอินพุตไปยังกลุ่มข้อมูลเอาต์พุต ในขั้นตอนของการส่งข้อมูล จะทำการแปลงค่าจากฟังก์ชั่นตามความต้องการของปัญหา

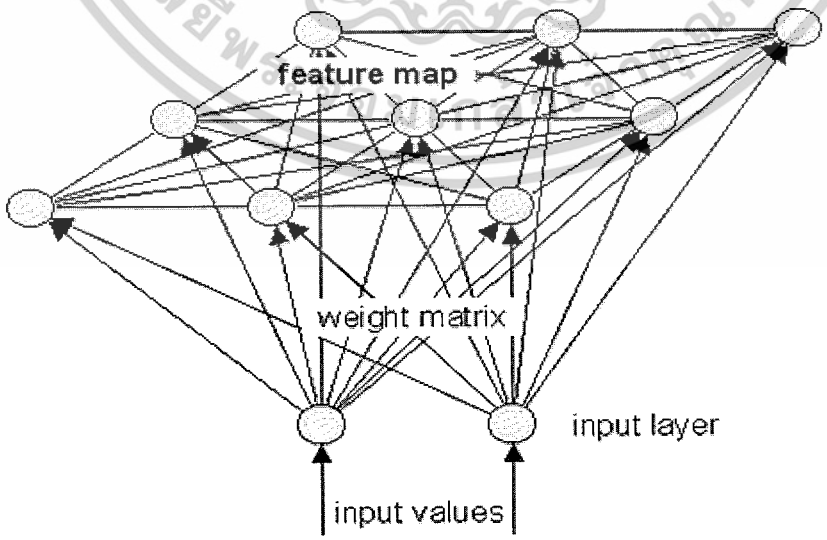
2. นิวรอนแบบส่งข้อมูลย้อนกลับ (Feedback Network) เป็นการส่งกลุ่มข้อมูลอินพุตไปยังกลุ่มข้อมูลเอาต์พุต แล้วนำข้อมูลย้อนกลับมาคำนวณใหม่

3. นิวรอนแบบแข่งขัน (Competitive Network) จะอาศัยข้อมูลข้างเคียงเพื่อเปรียบเทียบเซลล์ที่ใกล้เคียงกับข้อมูลอินพุตมากที่สุด ซึ่งเป็นการเรียนรู้แบบไม่ต้องอาศัยผู้สอน

2.4.1.5.2 หลักการทำงานของ Self Organizing Map (SOM)

SOM เป็นนิวรอนเน็ตเวิร์คแบบไม่มีผู้สอน วงจรข่ายจะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูล โครงสร้างโมเดลของ SOM ประกอบด้วยเซลล์ 2 ชั้น ชั้นแรกคือชั้นของอินพุต (Input Layer) ซึ่งประกอบด้วยเซตของอินพุตเวคเตอร์ 1 x n มิติ แทนด้วย x(t) ชั้นที่ 2 คือชั้นของเอาท์พุต ประกอบด้วยโหนดของนิวรอน ในแต่ละโหนดของนิวรอนจะประกอบด้วยเวคเตอร์น้ำหนัก แทนด้วย w_i

$$x(t) = \{x_1 , x_2 , x_3 , \dots , x_n\} \qquad (2.10)$$



รูปที่ 2.16 แบบแผนภาพจำลองโคโฮเนน

วิธีและหลักการทำงาน

1. ทำการกำหนด ค่าอัตราการเรียนรู้ (α) รัศมีของบริเวณโหนดใกล้เคียง (σ) จำนวนรอบในการเรียนรู้สูงสุด (T) และรู้น้ำหนักเริ่มต้นให้กับโหนดทุกโหนดโดยการสุ่ม
2. สุ่มเลือกอินพุตเวคเตอร์จากอินพุตทั้งหมด
3. เปรียบเทียบอินพุตเวคเตอร์กับโหนดทุกโหนดเพื่อหาโหนดชนะจากโหนดทั้งหมด โดยคำนวณจากสูตร

$$w_c = \min_i \| x_i(t) - w_i(t) \| \qquad (2.11)$$

เมื่อ w_c คือ โหนดของเวคเตอร์น้ำหนักที่ชนะ

$x_i(t)$ คือ อินพุตเวคเตอร์ในปัจจุบัน

$w_i(t)$ คือ เวคเตอร์น้ำหนักในปัจจุบัน

4. ทำการปรับเวคเตอร์น้ำหนักของโหนดชนะเพื่อให้เข้าใกล้อินพุตเวคเตอร์มากขึ้น และทำการปรับเวคเตอร์น้ำหนักของโหนดใกล้เคียงกับโหนดชนะ เพื่อให้อินพุตเวคเตอร์ถัดไปที่มีค่าใกล้เคียงมีโหนดชนะใหม่ใกล้เคียงกับโหนดชนะเดิม

โดยคำนวณจากสูตร

$$w_i(t+1) = w_i(t) + \alpha(t) h_{ci}(t)(x(t) - w_i(t)) \qquad (2.12)$$

เมื่อ t คือ รอบปัจจุบันของการเรียนรู้

$\alpha(t)$ คือ อัตราการเรียนรู้ในปัจจุบัน โดยอัตราการเรียนรู้นั้นขึ้นอยู่กับจำนวนรอบ แสดงเป็นสมการเชิงเส้นได้ดังนี้

$$\alpha(t) = \alpha(0)\left(\frac{T - t}{T}\right) \qquad (2.13)$$

เมื่อ T คือ จำนวนรอบทั้งหมด

$h_{ci}(t)$ คือ ฟังก์ชั่นที่ใช้ในการปรับค่าน้ำหนักของโหนดใกล้เคียง โดยทั่วไปจะใช้ฟังก์ชั่นเกาส์เซียน (Gaussian) ดังสมการ

$$h_{ci}(t) = \exp\left(\frac{-\| r_c - r_i \|^2}{2\sigma^2(t)}\right) \qquad (2.14)$$

เมื่อ $\| r_c - r_i \|$ คือ ระยะห่างจากโหนด i ถึงโหนดชนะ c

$\sigma(t)$ คือ รัศมีของบริเวณโหนดใกล้เคียง โดยการปรับรัศมีจะปรับดังสมการ

$$\sigma(t+1) = 1 + (\sigma(t)-1)\left(\frac{T - t}{T}\right) \qquad (2.15)$$

5. ทำซ้ำข้อ 2 จนครบทุกอินพุตเวคเตอร์
6. ทำซ้ำข้อ 2 จนครบจำนวนรอบในการเรียนรู้สูงสุด

บทที่ 3

ออกแบบและพัฒนาระบบ

3.1 เครื่องมือและภาษาโปรแกรมที่ใช้ในการพัฒนาระบบ

การพัฒนาระบบในโครงงานนี้ได้ใช้เครื่องมือและภาษาในการพัฒนา ดังนี้

3.1.1 ฮาร์ดแวร์ (Hardware)

เครื่องมือคอมพิวเตอร์ที่ใช้ในการพัฒนาระบบและทดสอบระบบ มีคุณสมบัติ ดังนี้

- Notebook Sony Vaio : Intel Core Duo Processer T2300 (1.66 GHz)
- RAM : 1 GB
- Hard Disk : 80 GB

3.1.2 ซอฟต์แวร์ (Software)

Software ที่ใช้ในการพัฒนาระบบและทดสอบระบบ

- Microsoft Window XP Professional(SP2)
- Java SE 5 SDK

3.1.3 เครื่องมือ (Tool)

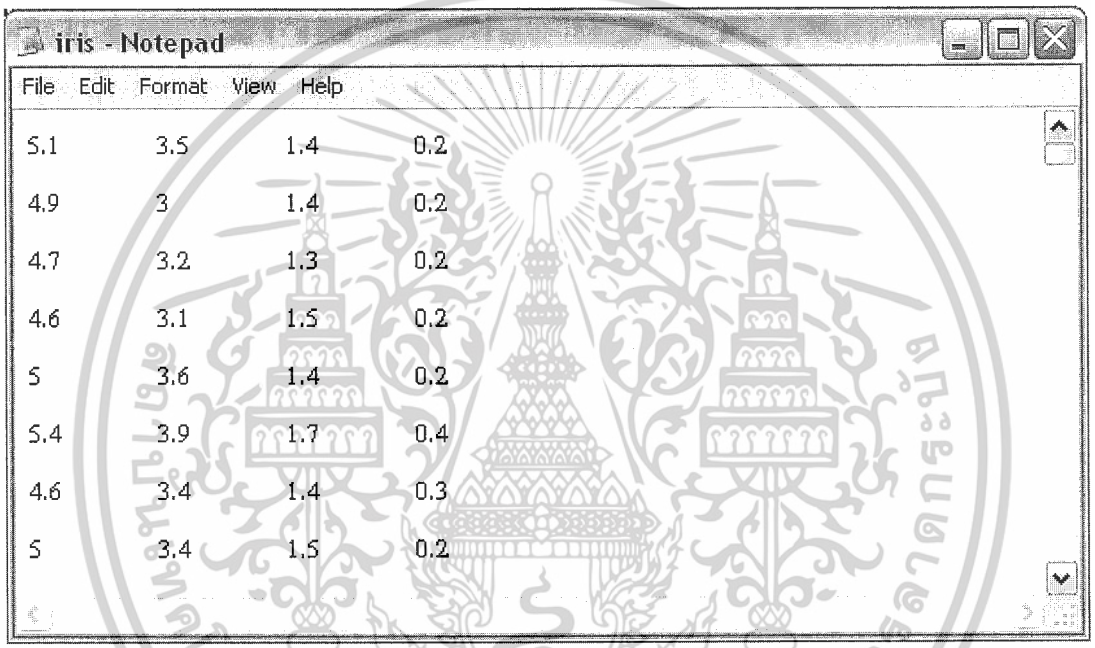
เครื่องมือที่ใช้ในการพัฒนาระบบและทดสอบระบบ มีคุณสมบัติ ดังนี้

- Programming Tool : NetBeans 5.5

3.2 อธิบายการทำงานของโปรแกรม

เป็นการอธิบายขั้นตอนการใช้งานโปรแกรม ซึ่งมีขั้นตอนดังนี้

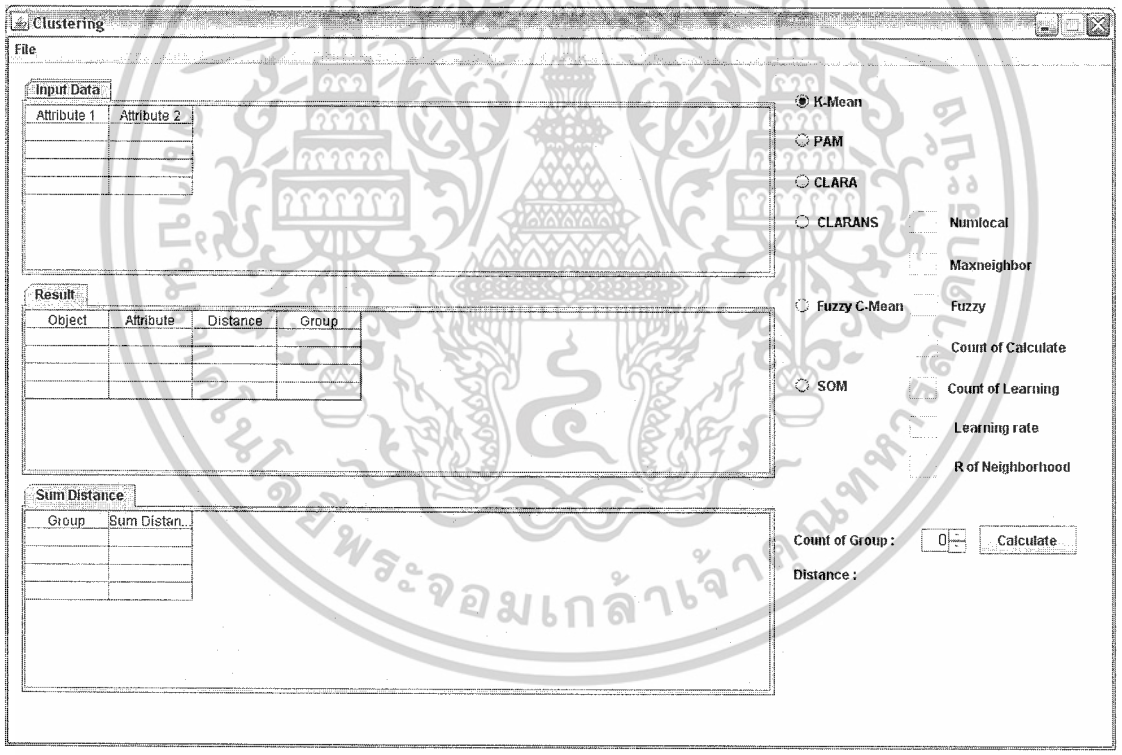
3.2.1 ตัวอย่าง Text file ข้อมูลที่ใช้อ่านข้อมูลเข้าสู่โปรแกรม โดยข้อมูลที่ใส่นั้นเป็นข้อมูลชนิดจำนวนเต็มที่ได้จากการแปลงข้อมูล ข้อมูล 1 บรรทัดแทน Object 1 Object ซึ่งประกอบด้วยหลาย Attribute โดยแต่ละ Attribute เว้นห่างกันด้วย Tab ดังรูปที่ 3.1 ซึ่งประกอบด้วย Object บรรทัดละ 1 Object แต่ละ Object ประกอบด้วย 4 Attribute



รูปที่ 3.1 ตัวอย่าง Text file

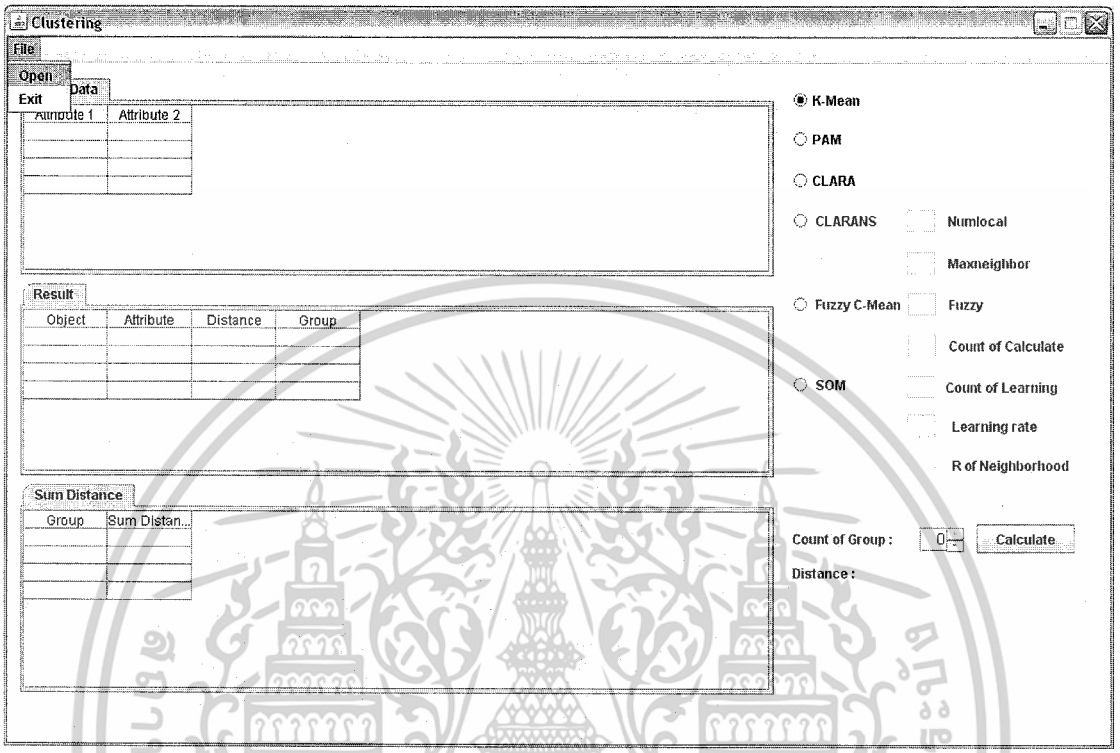
3.2.2 ตัวอย่างโปรแกรมที่พัฒนา ดังรูป 3.2 ซึ่งประกอบด้วย

- ตาราง Input Data จะแสดงข้อมูลที่จะนำมาคำนวณทั้งหมดจาก Text file
- ตาราง Result จะแสดงข้อมูลที่ได้จากการคำนวณซึ่งบ่งบอกว่า Object นั้นๆ มีค่า Distance ระหว่างแต่ละกลุ่มมีค่าเท่าใด และถูกจัดให้อยู่ในกลุ่มใด
- ตาราง Sum Distance แสดง จุดศูนย์กลางของกลุ่ม ผลรวม Distance ภายในกลุ่ม และจำนวนสมาชิกในกลุ่ม
- ปุ่มสำหรับให้เลือกใช้อัลกอลิทึม K-mean ,PAM, CLARA, CLARANS, Fuzzy C-Mean, Self Organizing Map (SOM)
- ช่อง Count of Group สำหรับใส่จำนวนกลุ่มที่ต้องการจะทำการแบ่ง
- ปุ่ม Calculate สำหรับทำการคำนวณ
- Label Distance สำหรับแสดงผลรวม Distance ของทุกกลุ่มที่ได้จากการคำนวณ



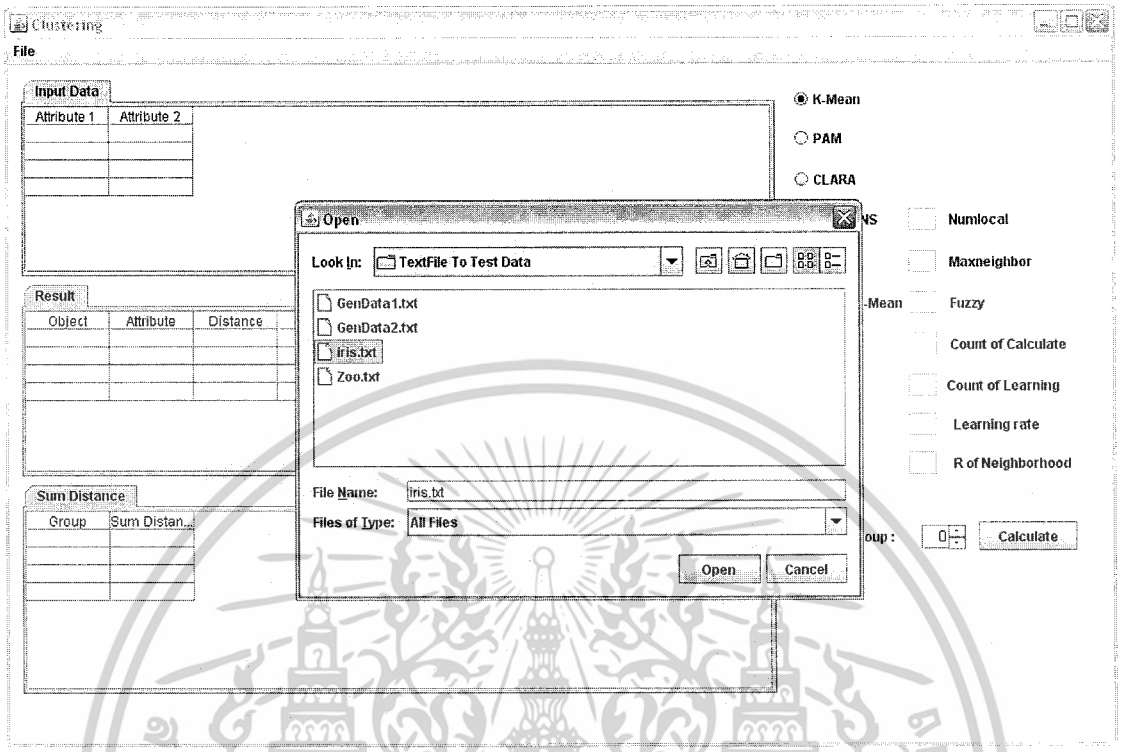
รูปที่ 3.2 ตัวอย่างหน้าจอโปรแกรม

3.2.3 เลือกที่ File และ Open ดังรูปที่ 3.3 เป็นการเปิด Dialog box เพื่อเลือก Text file ที่จะทำการอ่านข้อมูล



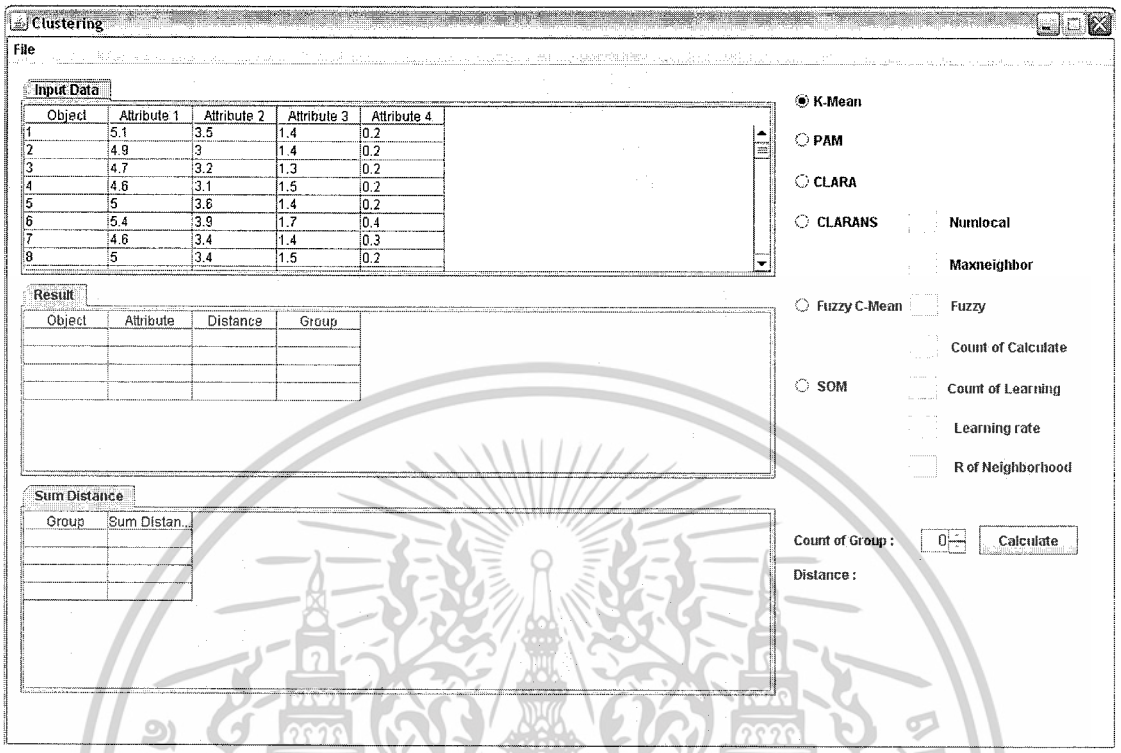
รูปที่ 3.3 ตัวอย่างการเปิด Dialog Box เพื่อเลือก Text file ของโปรแกรม

3.2.4 เลือก Text file ข้อมูลที่จะนำมาคำนวณ ดังรูป 3.4



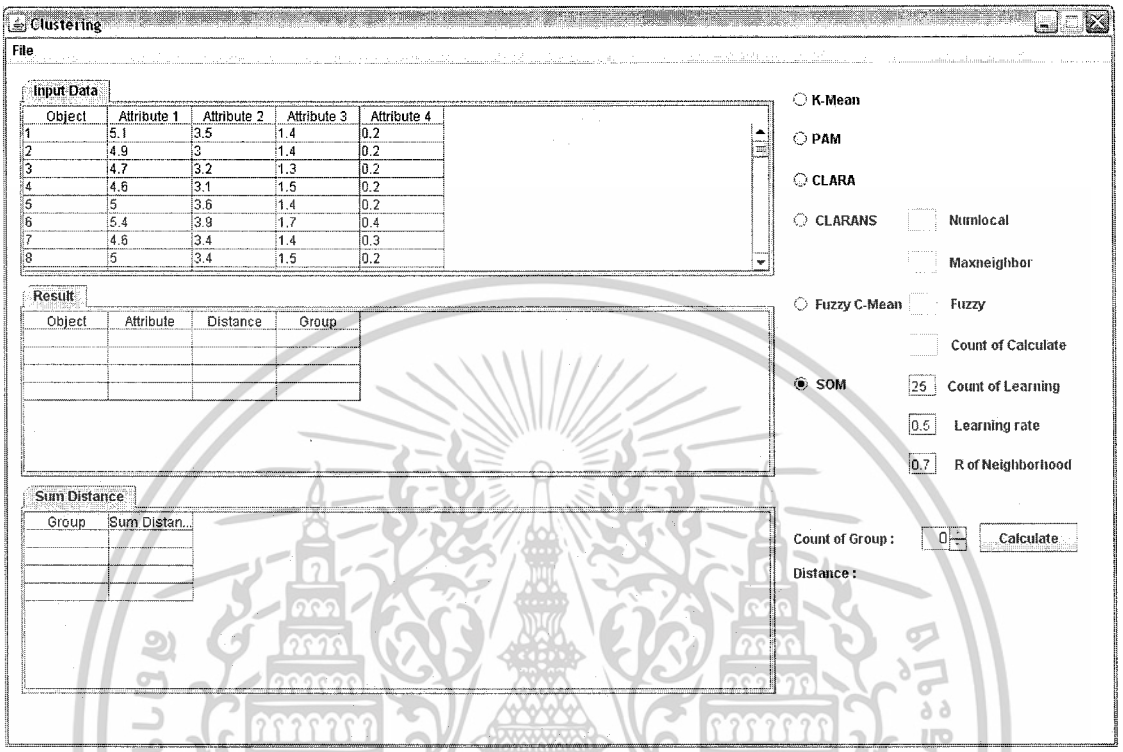
รูปที่ 3.4 ตัวอย่างการเลือก Text file ของโปรแกรม

3.2.5 ข้อมูลจาก Text file จะถูกแสดงที่ตาราง Input Data ดังรูป 3.5



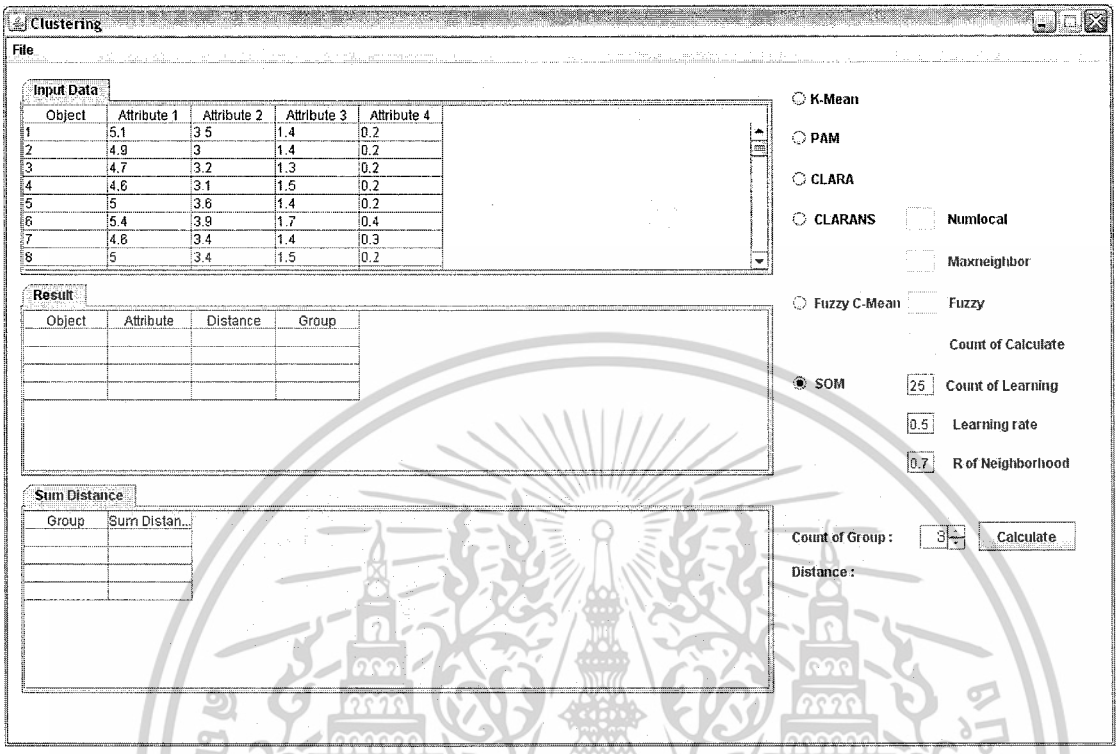
รูปที่ 3.5 ตัวอย่างการแสดงข้อมูล Input ของโปรแกรม

3.2.6 ทำการเลือกอัลกอลิทึมที่จะใช้ในการคำนวณ และใส่จำนวนกลุ่มที่ต้องการแบ่งในช่อง Count of Group โดยตัวอย่างนี้ใช้ SOM ในการคำนวณ ทำการใส่ค่าตัวแปร ดังรูป 3.6



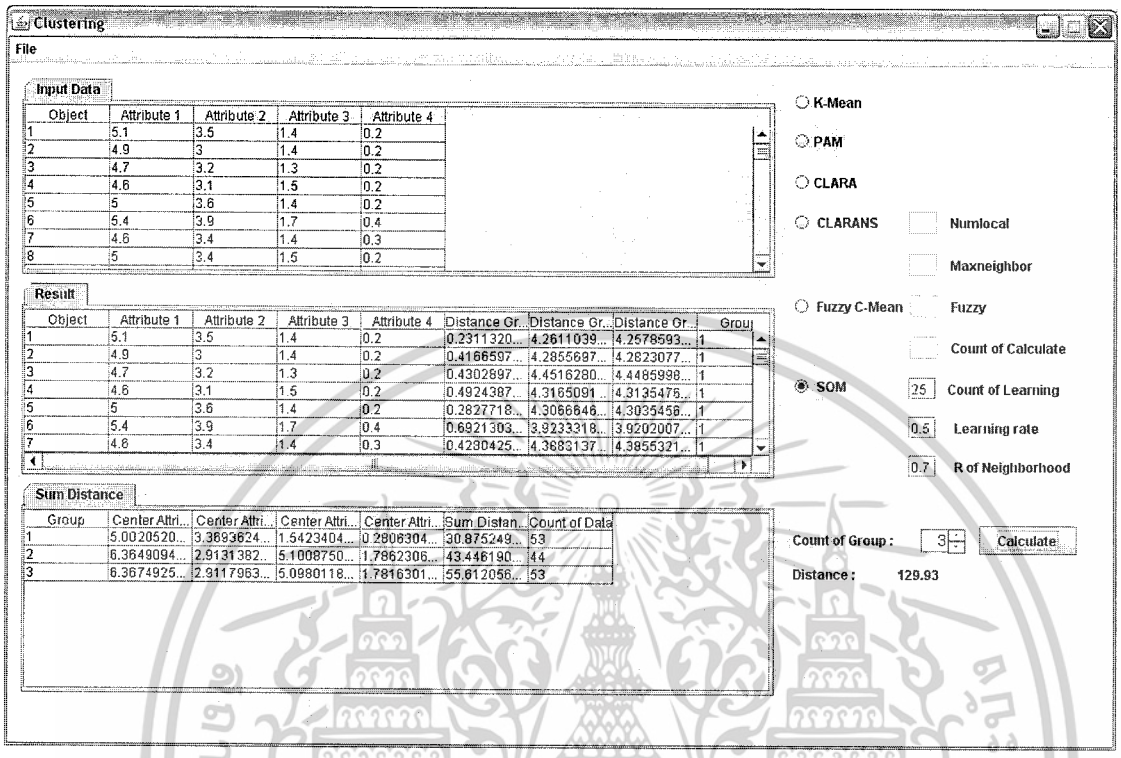
รูปที่ 3.6 ตัวอย่างการเลือกอัลกอริทึมของโปรแกรม

3.2.7 ทำการใส่ค่าจำนวนกลุ่มที่ต้องการจะทำการแบ่ง ดังรูปที่ 3.7



รูปที่ 3.7 ตัวอย่างการใส่ค่าข้อมูลของโปรแกรม

3.2.8 เมื่อกดปุ่ม Calculate โปรแกรมจะทำการคำนวณข้อมูล ตามเงื่อนไขที่ใส่ และผลลัพท์จากการคำนวณจะแสดงในตาราง Result ตาราง Sum Distance และ Label Distance ดังรูปที่ 3.8



รูปที่ 3.8 ตัวอย่างการแสดงผลการคำนวณของโปรแกรม

บทที่ 4

ผลการทดลอง

4.1 ข้อมูลที่ใช้ในการทดลอง

ในการทดลองนี้ ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลชุดของ Iris ซึ่งประกอบด้วยชุดข้อมูล Sepal length, Sepal width, Petal length, Petal width ทั้งหมด 150 ชุดข้อมูล และชนิดของข้อมูลเป็นตัวเลขทศนิยม

ตารางที่ 4.1 ตัวอย่างข้อมูลที่ใช้ในการทำการทดลอง

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5	3.6	1.4	0.2

4.2 การทดลอง

หลังจากการพัฒนาระบบ ได้ทำการทดสอบระบบโดยทำการเปรียบเทียบการจัดกลุ่มด้วยอัลกอริทึมต่างๆ ซึ่งมีรายละเอียดของผลการทดลองดังนี้

4.2.1 ทดลองโดยใช้ K-means algorithm

ทำการทดลองโดยแบ่งการทดลองออกเป็น 5 รอบ แต่ละรอบจะทำการสุ่มข้อมูลเพื่อเป็นจุดเริ่มต้นใหม่ทุกครั้ง ได้ผลการทดลองดังตารางที่ 4.2

ตารางที่ 4.2 ผลการทดลองโดยใช้ K-means algorithm

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.006	3.418	1.464	0.244	5.902	2.748	4.393	1.433	6.85	3.073	5.742	2.071	97.33	1.25
2	5.006	3.418	1.464	0.244	5.883	2.741	4.388	1.434	6.853	3.076	5.715	2.053	97.35	1.688
3	5.006	3.418	1.464	0.244	5.902	2.748	4.393	1.433	6.85	3.073	5.742	2.071	97.33	1.297
4	5.006	3.418	1.464	0.244	5.883	2.741	4.388	1.434	6.853	3.076	5.715	2.053	97.35	1.442
5	5.006	3.418	1.464	0.244	5.8836	2.741	4.3885	1.4344	6.8538	3.0769	5.7154	2.0538	97.35	1.593

- ค่าเฉลี่ยของ Distance = 97.342
- ส่วนเบี่ยงเบนมาตรฐาน = 0.01095

4.2.2 ทดลองโดยใช้ Fuzzy C-means algorithm

ทำการทดลองโดยแบ่งการทดลองออกเป็น 5 รอบ แต่ละรอบจะทำการสุ่มข้อมูลเพื่อเป็นจุดเริ่มต้นใหม่ทุกครั้งเหมือนการทดลองของ K-means Algorithm ได้ผลดังตารางที่ 4.3

ตารางที่ 4.3 ผลการทดลองโดยใช้ Fuzzy C-means algorithm

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.003	3.401	1.401	0.252	5.917	2.771	4.396	1.411	6.777	3.054	5.657	2.061	96.82	13.864
2	5.003	3.403	1.484	0.251	5.887	2.760	4.361	1.395	6.772	3.051	5.643	2.052	96.94	13.859
3	5.003	3.402	1.485	0.251	5.893	2.762	4.37	1.400	6.779	3.053	5.652	2.055	96.92	12.903
4	5.003	3.403	1.484	0.251	5.887	2.760	4.361	1.396	6.773	3.051	5.644	2.052	96.94	12.346
5	5.003	3.403	1.484	0.251	5.888	2.760	4.362	1.396	6.774	3.052	5.645	2.053	96.93	11.579

- ค่าเฉลี่ยของ Distance = 96.51
- ส่วนเบี่ยงเบนมาตรฐาน = 0.0509902

4.2.3 ทดลองโดยใช้ Self Organizing Map algorithm

ทำการทดลองโดยแบ่งการทดลองออกเป็น 5 รอบ แต่ละรอบจะทำการสุ่มข้อมูลเพื่อเป็นจุดเริ่มต้นใหม่ทุกครั้งเหมือนการทดลองของ K-means Algorithm และ Fuzzy C-means Algorithm และในการทดลอง Self Organizing Map Algorithm นี้จะทำการแบ่งการทดลองออกเป็น 3 ครั้ง โดยเริ่มต้นจะกำหนดให้การทดลองทั้ง 3 ครั้งมีจำนวนรอบของการเรียนรู้มีค่าเท่ากับ 20 และอัตราการเรียนรู้มีค่าเท่ากับ 0.5 แต่ระยะระหว่างโหนดใกล้เคียงของการทดลองทั้ง 3 ครั้งเป็น 0.5, 0.7, 0.9 ตามลำดับ

กำหนดให้ จำนวนรอบ = 20, อัตราการเรียนรู้ = 0.5, ระยะระหว่างโหนดใกล้เคียง = 0.5

ตารางที่ 4.4 ผลการทดลองโดยใช้ Self Organizing Map algorithm (a)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	19.984
2	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	19.751
3	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	19.657
4	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	20.328
5	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	19.845

- ค่าเฉลี่ยของ Distance = 130.44
- ส่วนเบี่ยงเบนมาตรฐาน = 0

กำหนดให้ จำนวนรอบ = 20, อัตราการเรียนรู้ = 0.5, ระยะระหว่างโหนดใกล้เคียง = 0.7

ตารางที่ 4.5 ผลการทดลองโดยใช้ Self Organizing Map algorithm (b)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	19.593
2	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	20.938
3	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	21.063
4	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	20.542
5	5.001	3.364	1.549	0.284	6.378	2.918	5.117	1.794	6.374	2.914	5.121	1.798	130.44	20.865

- ค่าเฉลี่ยของ Distance = 130.44
- ส่วนเบี่ยงเบนมาตรฐาน = 0

กำหนดให้ จำนวนรอบ = 20, อัตราการเรียนรู้ = 0.5, ระยะระหว่างโหนดใกล้เคียง = 0.9

ตารางที่ 4.6 ผลการทดลองโดยใช้ Self Organizing Map algorithm (c)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.001	3.364	1.549	0.284	6.378	2.914	5.118	1.794	6.375	2.918	5.120	1.798	130.45	19.469
2	5.001	3.364	1.549	0.284	6.378	2.914	5.118	1.794	6.375	2.918	5.120	1.798	130.45	19.985
3	5.001	3.364	1.549	0.284	6.378	2.914	5.118	1.794	6.375	2.918	5.120	1.798	130.45	19.891
4	5.001	3.364	1.549	0.284	6.378	2.914	5.118	1.794	6.375	2.918	5.120	1.798	130.45	20.241
5	5.001	3.364	1.549	0.284	6.378	2.914	5.118	1.794	6.375	2.918	5.120	1.798	130.45	19.956

- ค่าเฉลี่ยของ Distance = 130.45
- ส่วนเบี่ยงเบนมาตรฐาน = 0

4.2.4 ทดลองโดยใช้ PAM algorithm

ทำการทดลองโดยแบ่งการทดลองออกเป็น 3 รอบ แต่ละรอบจะทำการสุ่มข้อมูลเพื่อเป็นจุดเริ่มต้นใหม่ทุกครั้ง ได้ผลการทดลองดังตารางที่ 4.7

ตารางที่ 4.7 ผลการทดลองโดยใช้ PAM algorithm

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.0	3.4	1.5	0.2	6.0	2.9	4.5	1.5	6.8	3.0	5.5	2.1	98.21	46.375
2	5.0	3.4	1.5	0.2	6.0	2.9	4.5	1.5	6.8	3.0	5.5	2.1	98.21	46.109
3	5.0	3.4	1.5	0.2	6.0	2.9	4.5	1.5	6.8	3.0	5.5	2.1	98.21	45.734

- ค่าเฉลี่ยของ Distance = 98.21
- ส่วนเบี่ยงเบนมาตรฐาน = 0

4.2.5 ทดลองโดยใช้ CLARA algorithm

ทำการทดลองโดยแบ่งการทดลองออกเป็น 3 ครั้งโดยกำหนดจำนวนของข้อมูลในการสุ่มแต่ละครั้งมีค่าเท่ากับ 30, 45, 60 และทำการคำนวณแต่ละครั้ง 10 รอบ แต่ละรอบจะทำการสุ่มข้อมูลเพื่อเป็นจุดเริ่มต้นใหม่ทุกครั้ง

กำหนดให้ จำนวนข้อมูลในการสุ่มแต่ละครั้ง = 30

ตารางที่ 4.8 ผลการทดลองโดยใช้ CLARA algorithm (a)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.6	2.9	3.6	1.3	5.4	3.9	1.3	0.4	4.8	3.0	1.4	0.1	100.61	9.281
2	5.3	3.7	1.5	0.2	5.1	3.8	1.5	0.3	6.3	2.8	5.1	1.5	104.47	9.024
3	4.8	3.1	1.6	0.2	5.7	2.6	3.5	1.0	5.8	2.6	4.0	1.2	101.20	9.234
4	4.4	2.9	1.4	0.2	5.5	3.5	1.3	0.2	6.7	3.1	4.4	1.4	100.50	9.406
5	5.7	4.4	1.5	0.4	5.4	3.7	1.5	0.2	5.7	2.8	4.5	1.3	105.84	9.266
6	7.7	3.8	6.7	2.2	6.3	3.3	6.0	2.5	6.8	3.2	5.9	2.3	101.91	9.781
7	6.7	3.1	5.6	2.4	6.4	2.9	4.3	1.3	4.4	2.9	1.4	0.2	107.64	9.172
8	6.6	3.0	4.4	1.4	7.3	2.9	6.3	1.8	4.6	3.2	1.4	0.2	108.37	9.562
9	7.2	3.0	5.8	1.6	6.5	3.0	5.5	1.8	5.9	3.2	4.8	1.8	102.52	9.813
10	6.3	3.3	6.0	2.5	6.0	2.2	4.0	1.0	6.7	3.3	5.7	2.1	99.70	10.953

- ค่าเฉลี่ยของ Distance = 103.276
- ส่วนเบี่ยงเบนมาตรฐาน = 3.115228545

กำหนดให้ จำนวนข้อมูลในการสุ่มแต่ละครั้ง = 45

ตารางที่ 4.9 ผลการทดลองโดยใช้ CLARA algorithm (b)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.7	2.8	4.5	1.3	6.1	2.8	4.0	1.3	6.7	3.3	5.7	2.1	103.43	21.015
2	6.7	3.1	4.7	1.5	5.7	2.5	5.0	2.0	6.3	2.7	4.9	1.8	102.77	22.515
3	5.6	2.8	4.9	2.0	6.1	2.6	5.6	1.4	6.3	2.7	4.9	1.8	101.36	21.718
4	5.1	3.3	1.7	0.5	4.8	3.0	1.4	0.3	6.0	2.7	5.1	1.6	100.20	22.640
5	4.9	3.0	1.4	0.2	6.9	3.1	4.9	1.5	6.4	2.7	5.3	1.9	101.65	22.015
6	5.1	2.5	3.0	1.1	6.1	2.8	4.0	1.3	6.5	2.8	4.6	1.5	99.65	21.531
7	6.8	3.0	5.5	2.1	6.7	3.0	5.0	1.7	6.0	2.2	5.0	1.5	102.22	21.859
8	7.7	2.6	6.9	2.3	4.4	3.0	1.3	0.2	4.4	3.2	1.3	0.2	98.95	21.140
9	7.7	3.0	6.1	2.3	6.0	2.2	5.0	1.5	4.9	2.5	4.5	1.7	99.62	21.516
10	6.8	2.8	4.8	1.4	5.1	3.4	1.5	0.2	6.3	3.3	6.0	2.5	103.24	20.640

- ค่าเฉลี่ยของ Distance = 101.309
- ส่วนเบี่ยงเบนมาตรฐาน = 1.623113948

กำหนดให้ จำนวนข้อมูลในการสุ่มแต่ละครั้ง = 60

ตารางที่ 4.10 ผลการทดลองโดยใช้ CLARA algorithm (c)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.7	3.8	1.7	0.3	4.9	2.4	3.3	1.0	4.3	3.0	1.1	0.1	99.00	37.406
2	5.9	3.0	4.2	1.5	6.0	2.9	4.5	1.5	6.7	3.3	5.7	2.1	99.40	37.203
3	5.3	3.7	1.5	0.2	6.3	2.8	5.1	1.5	6.7	2.5	5.8	1.8	98.21	37.016
4	5.0	3.4	1.6	0.4	6.9	3.1	4.9	1.5	5.2	3.5	1.5	0.2	99.00	37.875
5	4.6	3.2	1.4	0.2	5.2	3.5	1.5	0.2	5.4	3.9	1.7	0.4	102.05	37.719
6	6.5	2.8	4.6	1.5	5.6	3.0	4.1	1.3	5.8	4.0	1.2	0.2	100.21	37.891
7	4.9	2.4	3.3	1.0	5.9	3.0	4.2	1.5	6.3	2.9	5.6	1.8	99.48	38.578
8	5.7	3.8	1.7	0.3	5.6	3.0	4.5	1.5	6.9	3.2	5.7	2.3	98.87	36.937
9	5.6	2.7	4.2	1.3	4.9	3.1	1.5	0.1	6.4	2.7	5.3	1.9	99.80	36.578
10	7.6	3.0	6.6	2.1	6.6	3.0	4.4	1.4	5.1	3.4	1.5	0.2	99.62	37.734

- ค่าเฉลี่ยของ Distance = 99.564
- ส่วนเบี่ยงเบนมาตรฐาน = 1.035022276

4.2.6 ทดลองโดยใช้ CLARANS algorithm

ทำการทดลอง โดยแบ่งการทดลองออกเป็น 3 ครั้ง โดยกำหนดค่า Numlocal มีค่าเท่ากับ 30, 45, 60 และ Maxneighbor มีค่าคงที่เท่ากับ 5 ทำการคำนวณแต่ละครั้ง 10 รอบ แต่ละรอบจะทำการสุ่มข้อมูลเพื่อเป็นจุดเริ่มต้นใหม่ทุกครั้ง

กำหนดให้ Numlocal = 30, Maxneighbor = 5

ตารางที่ 4.11 ผลการทดลองโดยใช้ CLARANS algorithm (a)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.1	3.5	1.4	0.2	6.2	2.9	4.3	1.3	6.3	3.3	6.0	2.5	112.83	24.204
2	5.1	3.5	1.4	0.2	5.9	3.0	4.2	1.5	6.3	3.3	6.0	2.5	113.14	24.781
3	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	24.688
4	5.1	3.5	1.4	0.2	6.1	3.0	4.6	1.4	6.3	3.3	6.0	2.5	111.19	24.328
5	5.1	3.5	1.4	0.2	5.7	2.9	4.2	1.3	6.3	3.3	6.0	2.5	113.20	24.515
6	4.8	3.4	1.6	0.2	5.7	2.9	4.2	1.3	6.3	3.3	6.0	2.5	115.02	25.718
7	5.1	3.5	1.4	0.2	6.1	3.0	4.6	1.4	6.3	3.3	6.0	2.5	111.19	24.063
8	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	24.563
9	5.0	3.4	1.5	0.2	5.7	2.9	4.2	1.3	6.3	3.3	6.0	2.5	112.29	24.328
10	5.1	3.5	1.4	0.2	6.1	2.8	4.7	1.2	6.3	3.3	6.0	2.5	113.90	24.282

- ค่าเฉลี่ยของ Distance = 112.108
- ส่วนเบี่ยงเบนมาตรฐาน = 1.931313888

กำหนดให้ Numlocal = 45, Maxneighbor = 5

ตารางที่ 4.12 ผลการทดลองโดยใช้ CLARANS algorithm (b)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.1	3.5	1.4	0.2	6.1	3.0	4.6	1.4	6.3	3.0	4.6	1.4	111.19	37.781
2	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	38.328
3	5.0	3.4	1.5	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	108.24	37.796
4	5.1	3.5	1.4	0.2	6.0	2.9	4.7	1.4	6.3	3.3	6.0	2.5	111.04	38.531
5	5.1	3.5	1.4	0.2	6.2	2.9	4.3	1.3	6.3	3.3	6.0	2.5	112.83	37.094
6	5.1	3.5	1.4	0.2	5.9	3.0	4.2	1.5	6.3	3.3	6.0	2.5	113.14	38.562
7	5.1	3.5	1.4	0.2	6.1	3.0	4.6	1.4	6.3	3.3	6.0	2.5	111.19	37.750
8	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	36.204
9	5.1	3.5	1.4	0.2	5.9	3.0	4.2	1.5	6.3	3.3	6.0	2.5	113.14	36.469
10	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	36.844

- ค่าเฉลี่ยของ Distance = 110.825
- ส่วนเบี่ยงเบนมาตรฐาน = 1.827264197

กำหนดให้ Numlocal = 60, Maxneighbor = 5

ตารางที่ 4.13 ผลการทดลองโดยใช้ CLARANS algorithm (c)

รอบ	Center Group 1				Center Group 2				Center Group 3				Distance	Time
1	5.1	3.5	1.4	0.2	5.7	3.5	1.4	1.3	6.3	3.3	6.0	2.5	112.06	50.25
2	5.0	3.4	1.5	0.2	6.1	2.9	4.7	1.4	6.3	3.3	6.0	2.5	110.06	50.469
3	5.1	3.5	1.4	0.2	6.1	3.0	4.6	1.4	6.3	3.3	6.0	2.5	111.19	50.688
4	5.1	3.4	1.5	0.2	6.1	3.0	4.6	1.4	6.3	3.3	6.0	2.5	110.67	48.547
5	5.1	3.4	1.5	0.2	6.1	3.0	4.6	1.4	6.3	3.3	6.0	2.5	110.67	48.813
6	5.1	3.5	1.4	0.2	5.9	3.0	4.2	1.5	6.3	3.3	6.0	2.5	113.14	54.312
7	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	50.500
8	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	52.438
9	5.1	3.5	1.4	0.2	6.2	2.9	4.3	1.3	6.3	3.3	6.0	2.5	112.83	62.016
10	5.1	3.5	1.4	0.2	6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	109.16	48.266

- ค่าเฉลี่ยของ Distance = 110.81
- ส่วนเบี่ยงเบนมาตรฐาน = 1.490167776

4.3 วิเคราะห์ผลการทดลอง

จากการดำเนินการศึกษาและทดลอง ได้กำหนดการสุ่มข้อมูลเริ่มต้นของแต่ละอัลกอริทึมมีค่าเท่ากัน และได้ผลสรุปดังนี้

1. การจัดกลุ่มโดยใช้ K-means algorithm มีการคำนวณที่รวดเร็ว แต่ประสิทธิภาพของการจัดกลุ่มจะขึ้นอยู่กับข้อมูลที่สุ่มขึ้นมาเป็นตัวแทนตอนแรก ถ้าสุ่มได้ข้อมูลที่ไม่ดี ก็จะได้การจัดกลุ่มที่ไม่มีประสิทธิภาพ

2. การจัดกลุ่มโดยใช้ Fuzzy C-means Algorithm มีประสิทธิภาพมากในการจัดกลุ่มกับข้อมูลที่มีความคลุมเคลือสูง เนื่องจากการจัดกลุ่มมีการคำนวนค่าความเป็นสมาชิกของข้อมูลเปรียบเทียบในกลุ่มต่างๆ

3. จากผลลัพธ์ที่ได้จากการคำนวนโดยใช้ Self Organizing Map Algorithm

ตารางที่ 4.16 ผลการทดลองโดยใช้ Self Organizing Map algorithm

Distance of Neighborhood	AVG	SD
0.5	130.44	0
0.7	130.44	0
0.9	130.45	0

- จากผลการทดลองปรากฏว่า ผลของการคำนวณในการจัดกลุ่มไม่ว่าจะทำการเปลี่ยนค่าเริ่มต้นของการจัดกลุ่มแต่ละครั้งของการคำนวณอย่างไร ผลลัพธ์ที่ได้จะใกล้เคียงกันมาก แม้ว่าจะทำการเปลี่ยนแปลงค่าระยะห่างระหว่างโหนดใกล้เคียง ก็ได้ผลลัพธ์เหมือนเดิม

4. การจัดกลุ่มโดยใช้ PAM algorithm เวลาที่ใช้ในการประมวลผลข้อมูล ขึ้นอยู่กับจำนวนของข้อมูลที่นำมาแบ่งกลุ่ม ซึ่งมีจำนวนของข้อมูลยิ่งมากเท่าไหร่ เวลาที่ใช้ในการประมวลผลก็จะยิ่งมีมากขึ้น แต่การประมวณผล จะได้ผลลัพท์ที่มีประสิทธิภาพในการแบ่งกลุ่มมาก เนื่องจากข้อมูลทุกตัวถูกนำมาประมวณผลทั้งหมด

5. จากผลลัพธ์ที่ได้จากการคำนวนโดยใช้ CLARA Algorithm

ตารางที่ 4.14 ผลการทดลองโดยใช้ CLARA algorithm

Random	AVG	S.D.
30	103.276	3.115229
45	101.309	1.623114
60	99.564	1.035022

- เนื่องจากค่าเฉลี่ยของ Distance และส่วนเบี่ยงเบนมาตรฐานมีค่าลดลง เมื่อจำนวนของการสุ่มข้อมูลในการคำนวณมีค่าเพิ่มขึ้น ดังนั้นจึงสรุปได้ว่าการจัดกลุ่มโดยใช้ CLARA Algorithm นั้น ถ้ายิ่งเพิ่มจำนวนของการสุ่มข้อมูลมากขึ้น ผลลัพธ์ที่ได้จากการจัดกลุ่มก็จะยิ่งดีขึ้น แต่ระยะเวลาที่ใช้ในการคำนวณจะมากขึ้นตามจำนวนข้อมูลที่สุ่ม

6. จากผลลัพธ์ที่ได้จากการคำนวนโดยใช้ CLARANS Algorithm

ตารางที่ 4.15 ผลการทดลองโดยใช้ CLARANS algorithm

Numlocal	Maxneighbor	AVG	SD
30	5	112.108	1.931313888
45	5	110.825	1.827264197
60	5	110.81	1.490167776

- เนื่องจากค่าเฉลี่ยของ Distance และส่วนเบี่ยงเบนมาตรฐานมีค่าลดลง เมื่อค่าของจำนวนครั้งในการคำนวน (Numlocal) มีค่าเพิ่มขึ้น เมื่อจำนวนครั้งของการสลับ (Maxneighbor) คงที่ ดังนั้นจึงสรุปได้ว่าการคำนวนโดยใช้ CLARANS Algorithm นั้น ถ้าเพิ่มจำนวนครั้งในการคำนวนมากขึ้น จะทำให้ได้ผลลัพธ์ของการจัดกลุ่มข้อมูลที่ดีขึ้น แต่ระยะเวลาที่ใช้ในการคำนวณจะมากขึ้นตามจำนวนครั้งในการคำนวน

บทที่ 5
สรุปผลการทดลองและข้อเสนอแนะ

5.1 สรุปผลการทดลอง

จากการทดลองและทำการวิเคราะห์ผลลัพธ์ที่ได้ ทำให้สามารถสรุปข้อเปรียบเทียบของการจัดกลุ่มด้วยอัลกอริทึมต่างๆ ดังนี้

1. เปรียบเทียบการจัดกลุ่มโดยใช้ K-means algorithm และ Fuzzy C-means Algorithm การจัดกลุ่มของทั้ง 2 อัลกอลิทึมต่างก็ให้ผลลัพธ์ของการจัดกลุ่มที่ดี มีค่าเฉลี่ยของ Distance และค่าส่วนเบี่ยงเบนมาตรฐานมีค่าต่ำทั้งคู่ แต่ส่วนเบี่ยงเบนมาตรฐานของ Fuzzy C-means Algorithm มีค่ามากกว่า K-means algorithm หลายเท่า เนื่องจากข้อมูลที่นำมาเป็นตัวอย่างในการจัดกลุ่มเป็นข้อมูลที่ไม่มีความคลุมเคลือ ดังนั้น การจัดกลุ่มโดยใช้ K-means algorithm เหมาะกับข้อมูลที่ไม่มีความคลุมเคลือ ส่วนข้อมูลที่มีความคลุมเคลือควรใช้ Fuzzy C-means Algorithm

2. เปรียบเทียบการจัดกลุ่มโดยใช้ PAM algorithm, CLARA Algorithm และ CLARANS Algorithm จากการทดลอง PAM algorithm มีประสิทธิภาพในการจัดกลุ่มมากที่สุด แต่ถ้ามีข้อมูลจำนวนมาก จะทำให้ใช้เวลาในการประมวณผลนาน ส่วนการจัดกลุ่มโดยใช้ CLARA Algorithm และ CLARANS Algorithm ใช้เวลาในการประมวณผลรวดเร็วกว่า PAM algorithm แต่ประสิทธิภาพในการจัดกลุ่มจะต่ำกว่า ถ้ามีข้อมูลจำนวนมากๆ CLARANS Algorithm จะมีประสิทธิภาพในการจัดกลุ่มดีกว่า CLARA Algorithm

5.2 ข้อเสนอแนะ

1. ในการจัดกลุ่มข้อมูลด้วยอัลกอริทึมแบบต่างๆ ค่าพารามิเตอร์ที่กำหนดในแต่ละอัลกอริทึมมีผลทำให้การจัดกลุ่มมีประสิทธิภาพ ดังนั้น จึงควรปรับค่าพารามิเตอร์ให้มีช่วงที่กว้างมากขึ้นเพื่อให้ได้ผลลัพท์หลายๆแบบ ซึ่งจะทำให้ได้ผลลัพธ์ของการจัดกลุ่มที่ดีขึ้น

2. ในการทดลองเพื่อดูผลลัพธ์ของการจัดกลุ่มข้อมูล ควรดูลักษณะของข้อมูลที่จะนำมาใช้ว่าเป็นแบบใด เหมาะกับการคำนวณด้วยอัลกอริทึมใด และควรใช้ข้อมูลหลายๆชุด เพื่อให้มีข้อเปรียบเทียบในการจัดกลุ่มข้อมูล

บรรณานุกรม

Jiawei Han and Micheline Kamber. 2006. **Data Mining Concepts and Techniques Second Edition.** United States: Morgan Kaufmann Publishers is an imprint of Elsevier. 500 Sansome Street, Suite 400, San Francisco, CA 94111